MIP - Melhoramento da Identificação de Perdas - Manual do Usuário

Ufes - Universidade Federal do Espírito Santo

2 de dezembro de 2005

Sumário

1 Introdução **5**

2 Instalando o sistema **7**
2.1 Instalando o PostgreSQL 7
2.2 Configurando o PostgreSQL para inicialização manual 13
2.3 Iniciando e finalizando o PostgreSQL 14
2.4 Configurando as permissões do servidor e criando o banco de dados do MIP . . . 15
2.5 Executando o sistema 20

3 Módulo de Aprendizagem **21**
3.1 Conhecendo as interfaces do Módulo de Aprendizagem 21
3.1.1 Janela inicial 21
3.1.2 Menu Arquivo 21
3.1.3 Seleção da base de dados 23
3.1.4 Seleção e configuração do classificador 23
3.1.5 Seleção e configuração do estimador de erro 24
3.1.6 Fila de execução de tarefas 24
3.1.7 Gerenciador de Classificadores 25
3.1.8 Gerenciador de Buscas 26
3.2 Funcionalidades do Módulo de Aprendizagem 28
3.2.1 Criando uma tarefa de treinamento 28
3.2.2 Executando um treinamento 29
3.2.3 Agendando a execução do treinamento 30
3.2.4 Verificando resultados e exportando um classificador 31
3.2.5 Criando, recuperando e salvando configurações 33
3.2.6 Técnicas de classificação e estimativa de erro disponíveis e seus parâmetros configuráveis 34
3.2.7 Busca por parâmetros de uma técnica de classificação 40

4 Módulo de Consultas **43**
4.1 Conhecendo as interfaces do Módulo de Consultas 43
4.1.1 Janela inicial 43
4.1.2 Menu Arquivo 43
4.1.3 Seleção da base de dados 43
4.1.4 Manipulação e seleção de técnicas 45
4.1.5 Modos de avaliação 45
4.1.6 Botões de controle e barra de estado 46
4.1.7 Janela de resultados 46
4.2 Funcionalidades do Módulo de Consultas 46
4.2.1 Importando e removendo um classificador 46
4.2.2 Realizando uma consulta 48
4.2.3 Verificando resultados 50
4.2.4 Criando, recuperando e salvando configurações 51

5 Preparando e selecionando dados **55**
5.1 Selecionando bases de dados no sistema 55
5.2 Selecionando uma base de dados em formato CSV 55
5.3 Obtendo informações da base 58
5.4 Configurando a base de dados 58
5.4.1 Inserindo uma descrição para base 59
5.4.2 Gerenciando as colunas da base 60
5.4.3 Gerenciando os filtros de descarte 65
5.4.4 Salvando e recuperando configurações 68
5.4.5 Considerações sobre a configuração da base para consultas 71
5.5 Selecionando uma base em formato incompatível com a configuração atual do sistema 72
5.6 Selecionando uma base de dados em formato de saída do Datawarehouse 72

A Formato dos arquivos de entrada do conversor **77**

B Formato do arquivo CSV gerado pelo conversor **83**

C Notas sobre os procedimentos de conversão **87**

D Glossário **89**

Capítulo 1

Introdução

Este manual tem como finalidade descrever as funções básicas do sistema MIP - Melhoramento de Identificação de Perdas - desenvolvido por uma equipe da UFES em projeto de P&D da ESCELSA. O sistema é composto por dois módulos: um chamado Módulo de Aprendizagem (treinamento) e outro chamado Módulo de Consultas.

O Módulo de Aprendizagem possibilita o treinamento de classificadores que serão usados no Módulo de Consultas.

O Módulo de Consultas possibilita a classificação dos clientes em candidatos ou não à inspeção.

Neste manual, cada um dos módulos é descrito em detalhes.

Capítulo 2

Instalando o sistema

O sistema MIP tem como pré-requisito de instalação, um computador com máquina virtual Java 1.5 previamente instalada. Neste manual não será descrito o procedimento de instalação da máquina virtual Java 1.5. Maiores informações e o programa de instalação da máquina virtual podem ser obtidos em http://java.sun.com/.

O sistema MIP é compatível com os sistemas operacionais Linux e Windows. A instalação do sistema requer dois passos:

- Instalação do banco de dados PostgreSQL
- Cópia dos arquivos do sistema MIP

Na próxima seção é descrito o procedimento de instalação do PostgreSQL no sistema operacional Windows. Este passo de instalação só é necessário se o usuário for utilizar as funções de conversão de dados do Datawarehouse. Caso o usuário deseje instalar o MIP no linux, as instruções para instalação do PostgreSQL no Linux podem ser encontradas em http://www.postgresql.org/.

Os arquivos do MIP devem ser copiados do CD de instalação. Os arquivos encontram-se no diretório “Instalacao/Sistemas” do CD. Para executar o sistema, basta copiar o diretório MIP para alguma pasta do disco do computador e executá-lo conforme descrito na seção 2.5. A única restrição em relação à pasta de destino dos arquivos no computador, é que ela deve possuir permissão de escrita (esta restrição impossibilita que o sistema seja executado diretamente do CD).

2.1 Instalando o PostgreSQL

Nesta seção são mostrados os passos de instalação do sistema PostgreSQL. A instalação do PostgreSQL somente é necessária caso o usuário deseje converter arquivos que estejam no formato do Datawarehouse da ESCELSA para o formato de entrada do MIP. Caso o usuário tenha os arquivos já em formato apropriado para o sistema MIP (ver capítulo 5), esta instalação não é necessária.

O procedimento de instalação aqui descrito se aplica ao sistema operacional Windows, nas versões XP, 2000 ou 2003. Para instalação no sistema operacional Linux ou em outras versões do Windows, procure informações em http://www.postgresql.org/.

Para iniciar a instalação, execute o arquivo postgresql-8.0.msi que se encontra no diretório “Instalacao/Postgresql” do CD. Deverá ser mostrada a tela da figura 2.1.

Selecione o idioma da instalação (Português) e clique em “Start”. Na figura 2.2 é mostrada a segunda tela da instalação. O assistente sugere que todos os programas sejam fechados antes de

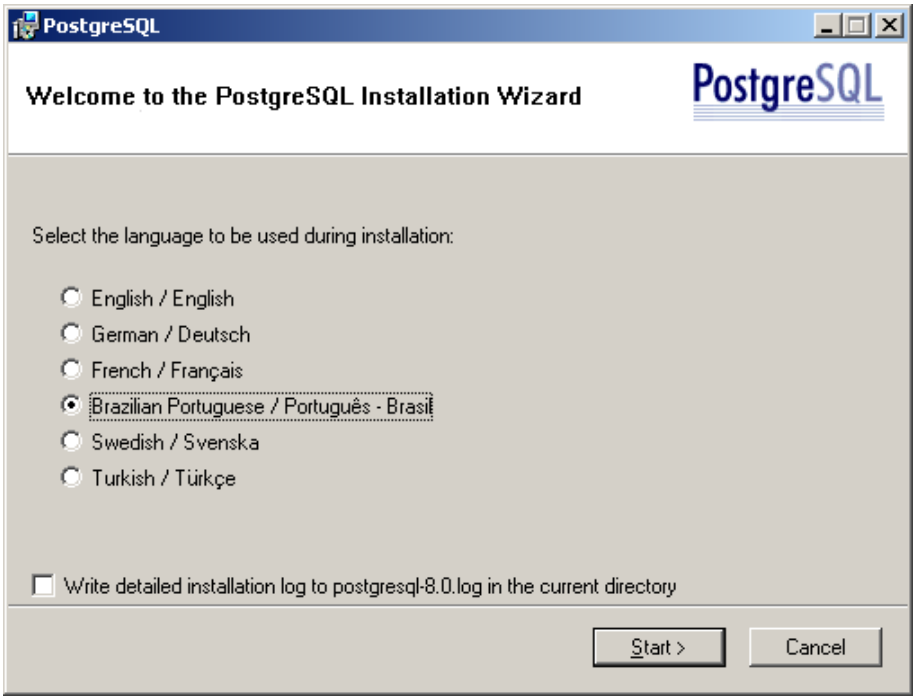

Figura 2.1: Instalação - tela de início

prosseguir. Feche outros programas abertos e clique em “Próximo”.

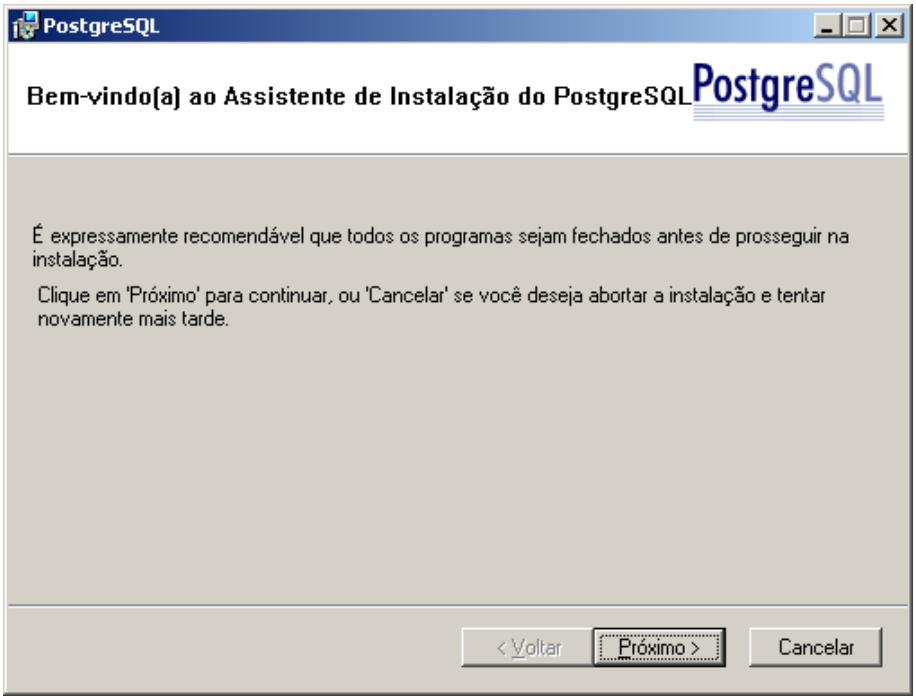
PostgreSQL
Bem-vindo(a) ao Assistente de Instalação do PostgreSQL
PostgreSQL
É expressamente recomendável que todos os programas sejam fechados antes de prosseguir na instalação.
Clique em 'Próximo' para continuar, ou 'Cancelar' se você deseja abortar a instalação e tentar novamente mais tarde.
< Voltar
Próximo >
Cancelar

Figura 2.2: Instalação - tela informativa

Após clicar em “Próximo”, será mostrada a tela da figura 2.3. Nesta tela é exibida apenas uma mensagem informativa. Clique novamente em “Próximo”.

Agora será exibida a tela da figura 2.4. Nesta tela são mostradas as opções de componentes disponíveis para instalação. Para uso do MIP bastam as opções padrões. Não altere nenhuma configuração e clique novamente em “Próximo”.

PostgreSQL
Informações de instalação
PostgreSQL
Por favor, leia todas as informações e instruções de instalação antes de prosseguir.
PostgreSQL 8.0
Informações de instalação
Bem-vindo ao Assistente de Instalação do PostgreSQL 8.0.
Antes de começar
Por favor, verifique a seção de FAQ em
http://pginstaller.projects.postgresql.org/FAQ_windows.html antes de começar.
Se você experimentar qualquer problema, verifique-a de novo. Além disso,
certifique-se de ter lido completamente essas informações de instalação.
< Voltar
Próximo >
Cancelar

Figura 2.3: Instalação - mensagem informativa

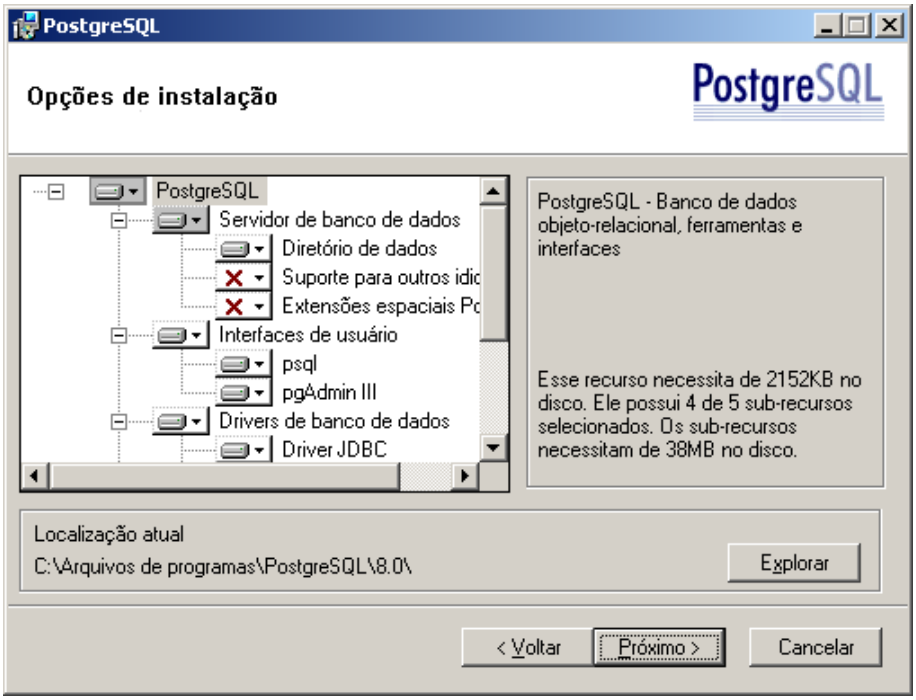

Figura 2.4: Instalação - seleção de componentes

Será mostrada a tela da figura 2.5. Nesta tela, é necessário preencher o nome de usuário que será criado na máquina na qual o sistema está sendo instalado. O nome de usuário sugerido pode ser mantido (postgres). Na caixa “Domínio” deve ser preenchido o nome do computador. Por padrão, esta caixa já conterá o nome correto.

Caso não seja preenchido o campo senha, o sistema criará uma senha aleatoriamente. Como esta senha não será mais necessária, recomenda-se que o campo não seja preenchido, para que o sistema crie uma senha automaticamente. Neste caso, será mostrada a tela de figura 2.6.

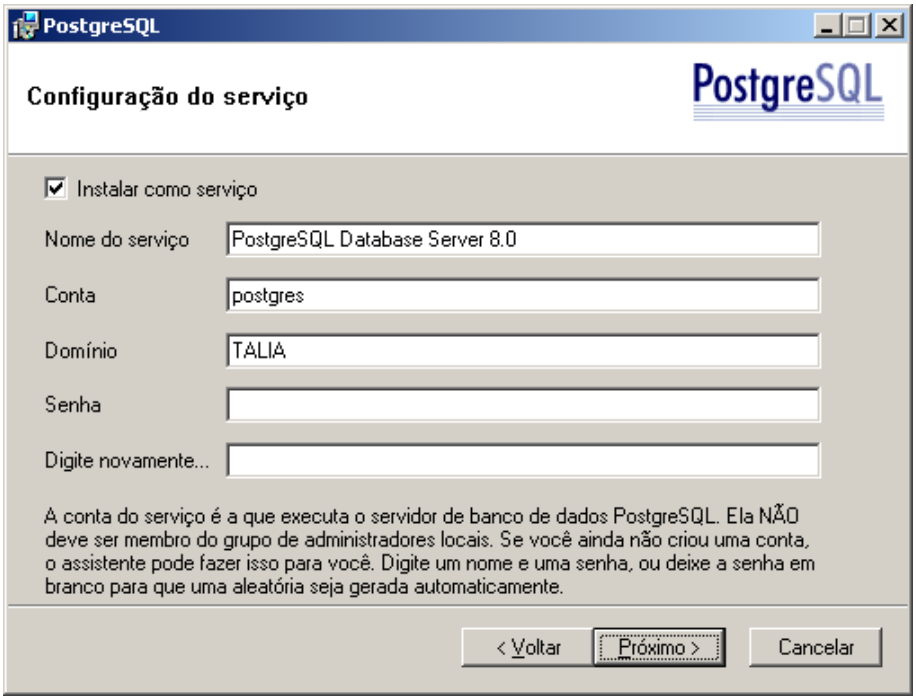
PostgreSQL
Configuração do serviço
PostgreSQL
Instalar como serviço
Nome do serviço
PostgreSQL Database Server 8.0
Conta
postgres
Domínio
TALIA
Senha
Digite novamente...
A conta do serviço é a que executa o servidor de banco de dados PostgreSQL. Ela NÃO deve ser membro do grupo de administradores locais. Se você ainda não criou uma conta, o assistente pode fazer isso para você. Digite um nome e uma senha, ou deixe a senha em branco para que uma aleatória seja gerada automaticamente.
< Voltar
Próximo >
Cancelar

Figura 2.5: Instalação - seleção de usuário e senha

Na figura 2.6 é mostrada a tela exibida após o acionamento do botão “Próximo”. Nesta tela é necessário clicar no botão “Sim”. Caso não seja digitada nenhuma senha na tela anterior (conforme recomendado), o sistema mostrará a tela da figura 2.7.

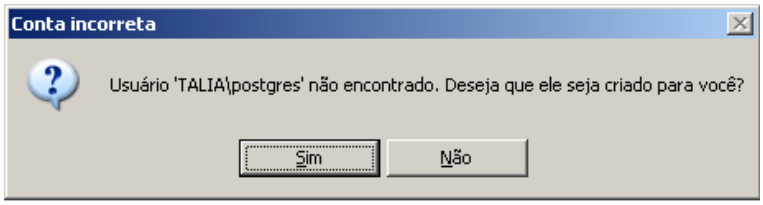

Figura 2.6: Instalação - confirmação da criação de usuário

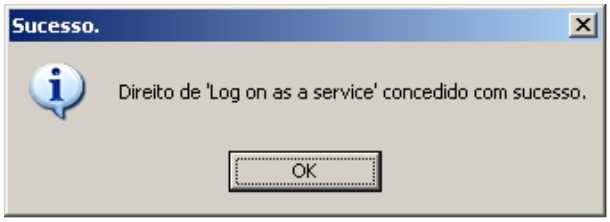

Figura 2.7: Instalação - o instalador cria uma senha aleatoriamente

A próxima tela (figura 2.8) contém opções importantes, que devem ser preenchidas com cuidado. Nela deverá ser criado o usuário administrador do banco de dados (não há relação alguma com o usuário do sistema operacional criado no passo anterior). Recomenda-se que o nome do usuário seja “postgres”. A senha, neste caso, deve ser digitada e não poderá ser esquecida. Depois de preencher todos os dados, clicar em “Próximo”.

Agora aparecerá a tela da figura 2.9. Nesta tela devem ser selecionadas as linguagens proce-

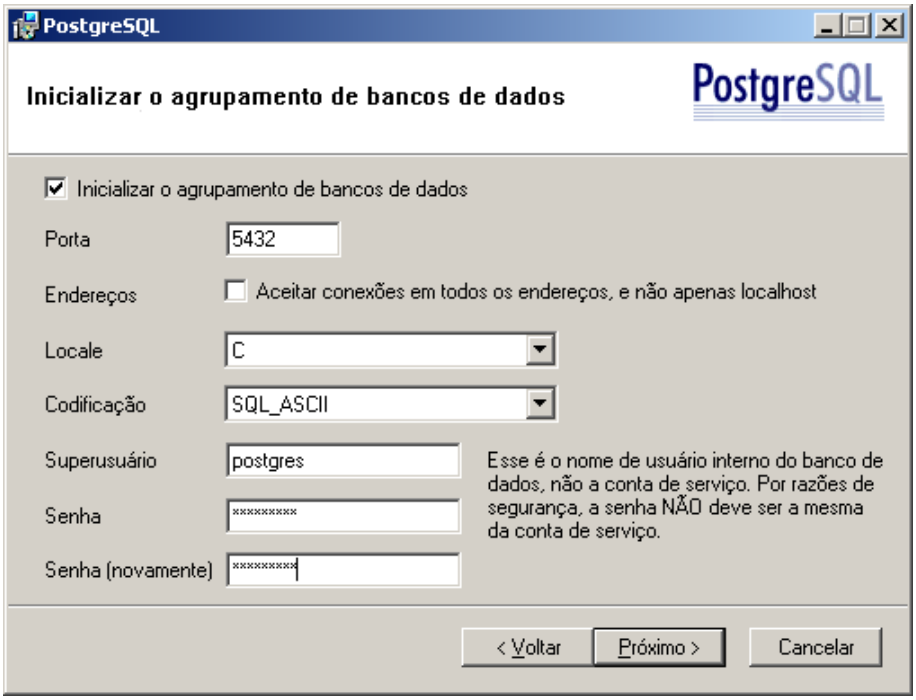

Figura 2.8: Instalação - criação do usuário administrador do sistema

durais a instalar. Deixar marcada somente a opção padrão (PL/pgsql) e clicar no botão “Próximo”.

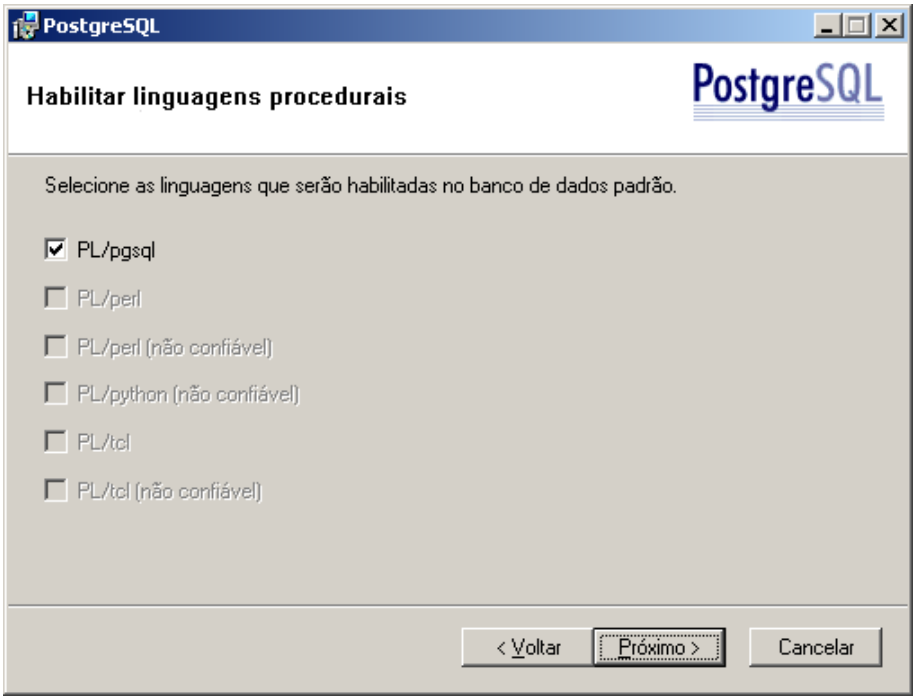

Figura 2.9: Instalação - seleção de linguagens procedurais

A próxima tela (figura 2.10) pede que sejam selecionados módulos adicionais a instalar. Novamente, nenhuma opção deve ser alterada. Clicar diretamente no botão “Próximo”.

O sistema está pronto para iniciar a instalação (figura 2.11). Clicar no botão “Próximo”. Durante a instalação será exibida uma barra de progresso conforme ilustrado na figura 2.12. Aguardar até que a instalação esteja concluída.

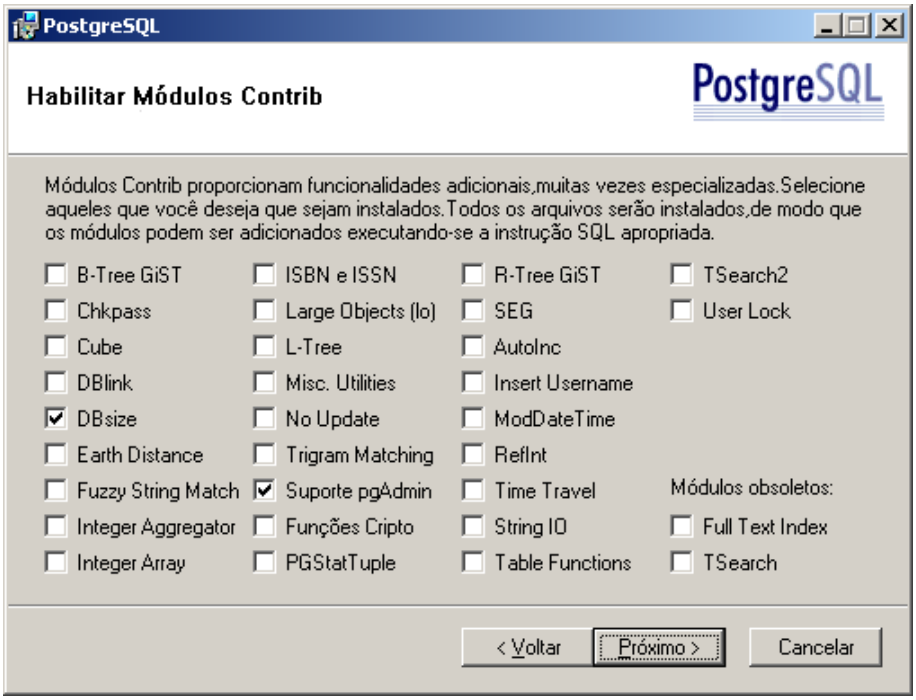

Figura 2.10: Instalação - seleção de módulos adicionais

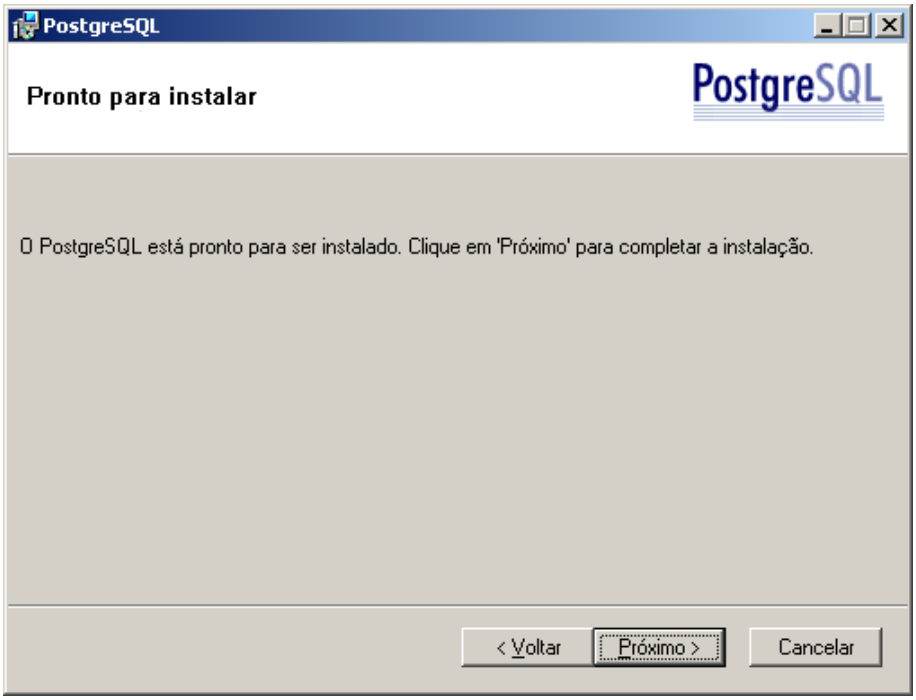

Figura 2.11: Sistema pronto para ser instalado

Ao terminar a instalação, será exibida a tela ilustrada na figura 2.13. Clicar no botão concluir. Agora o sistema está instalado. Porém ainda é necessário configurar o banco de dados para utilização do MIP.

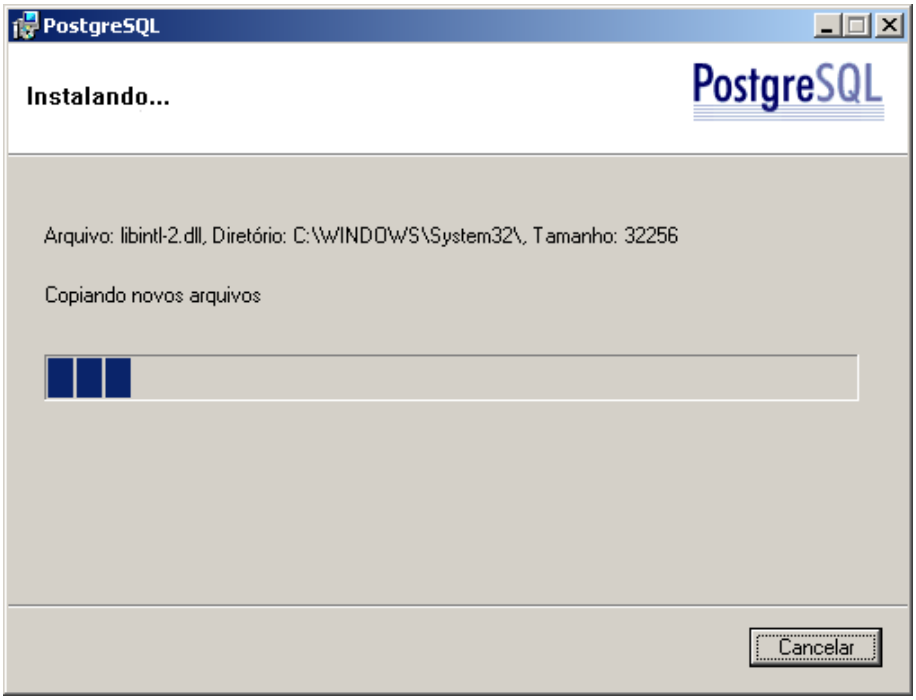

Figura 2.12: Barra de progresso durante a instalação

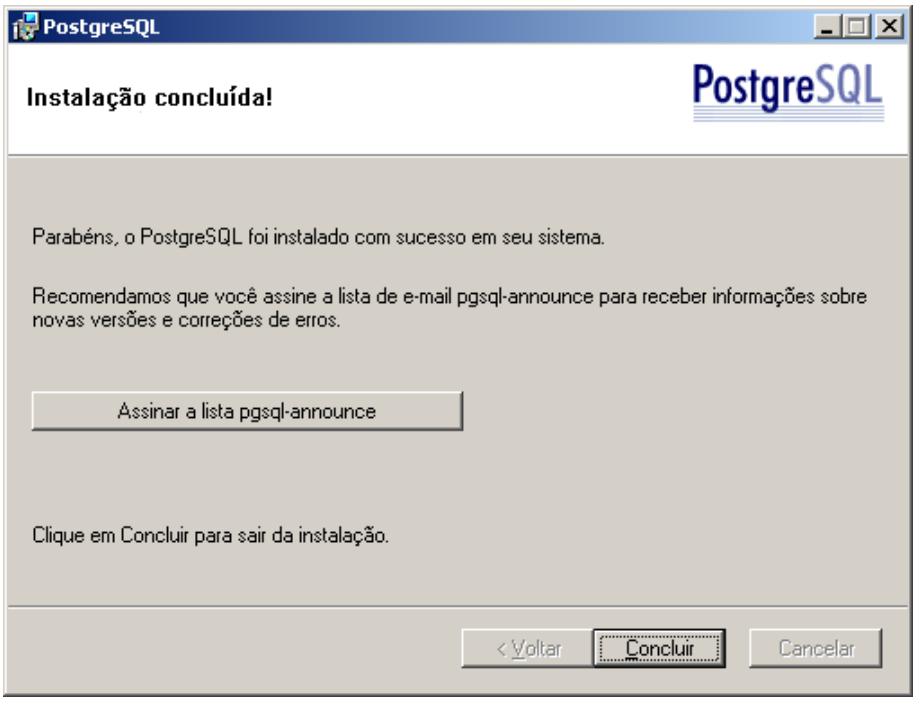
PostgreSQL
Instalação concluída!
PostgreSQL
Parabéns, o PostgreSQL foi instalado com sucesso em seu sistema.
Recomendamos que você assine a lista de e-mail pgsql-announce para receber informações sobre novas versões e correções de erros.
Assinar a lista pgsql-announce
Clique em Concluir para sair da instalação.
< Voltar
Concluir
Cancelar

Figura 2.13: Instalação concluída

2.2 Configurando o PostgreSQL para inicialização manual

O primeiro passo para configurar o sistema é desabilitar o início automático do banco de dados PostgreSQL. Por padrão, o sistema é instalado como serviço. Sendo um serviço, ele será iniciado automaticamente toda a vez que o computador for ligado. Porém isto não é necessário, dado que o banco de dados será usado esporadicamente. Para desabilitar o início automático do banco de dados, selecione a opção “Painel de Controle” no menu iniciar do Windows, conforme ilustrado

na figura 2.14.

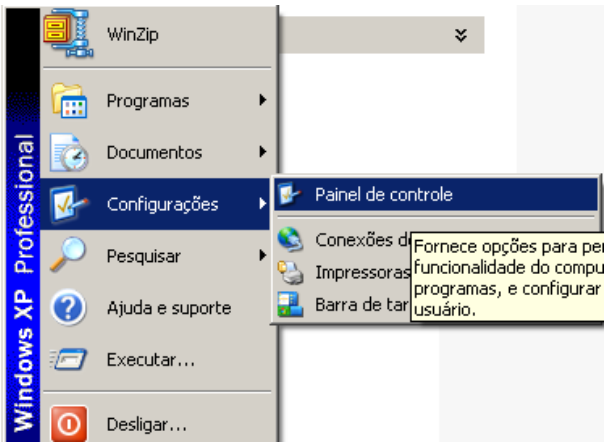

Figura 2.14: Configuração - seleção do “Painel de Controle” do Windows

Agora, localize na tela do “Painel de Controle” a opção “Ferramentas Administrativas”, conforme ilustrado na figura 2.15. Abra a opção “Ferramentas Administrativas” clicando duas vezes em seu nome.

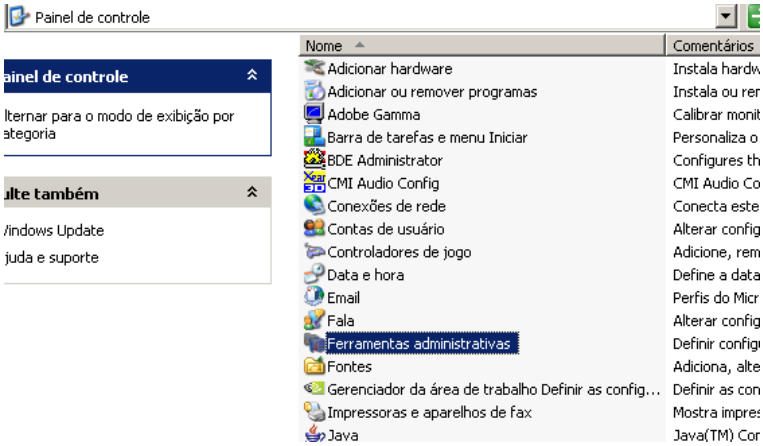

Figura 2.15: Configuração - seleção de “Ferramentas Administrativas”

Será aberta uma tela semelhante à ilustrada na figura 2.16. Clique duas vezes na opção “Serviços”.

Na tela de “Serviços”, localize a opção “PostgreSQL Database Server 8.0”, conforme ilustrado na figura 2.17. Clique duas vezes no serviço do PostgreSQL para abertura da tela de configuração.

Na figura 2.18 é ilustrada a tela de configuração do serviço do PostgreSQL. Alterar o tipo de inicialização de “Automático” para “Manual”. Clicar no botão “OK”.

2.3 Iniciando e finalizando o PostgreSQL

Caso o usuário tenha feito corretamente os passos anteriores (de configuração do serviço do PostgreSQL) será necessário iniciar manualmente o banco de dados toda vez que for preciso fazer uma conversão de dados do formato de saída do Datawarehouse para o formato de entrada do MIP.

Recomenda-se que o sistema PostgreSQL permaneça desligado enquanto o MIP não estiver sendo executado, para evitar o consumo desnecessário de recursos do computador.

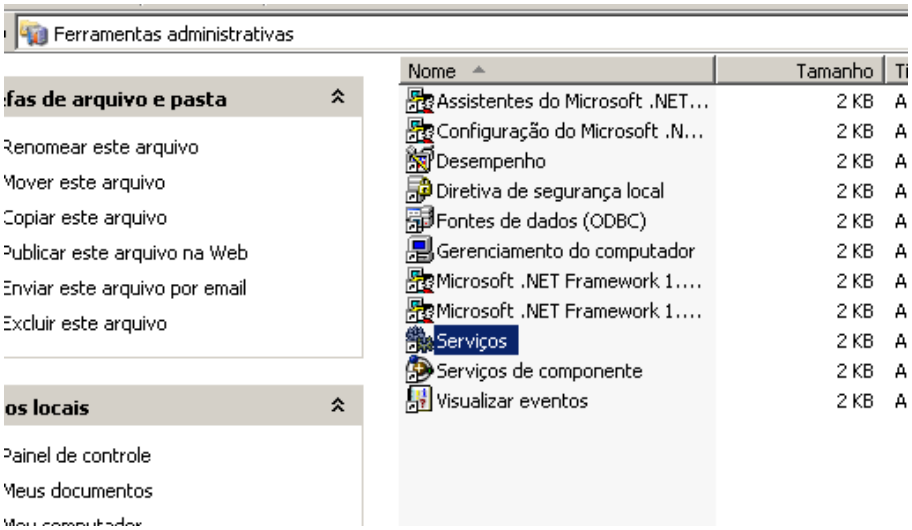

Figura 2.16: Configuração - seleção da opção “Serviços”

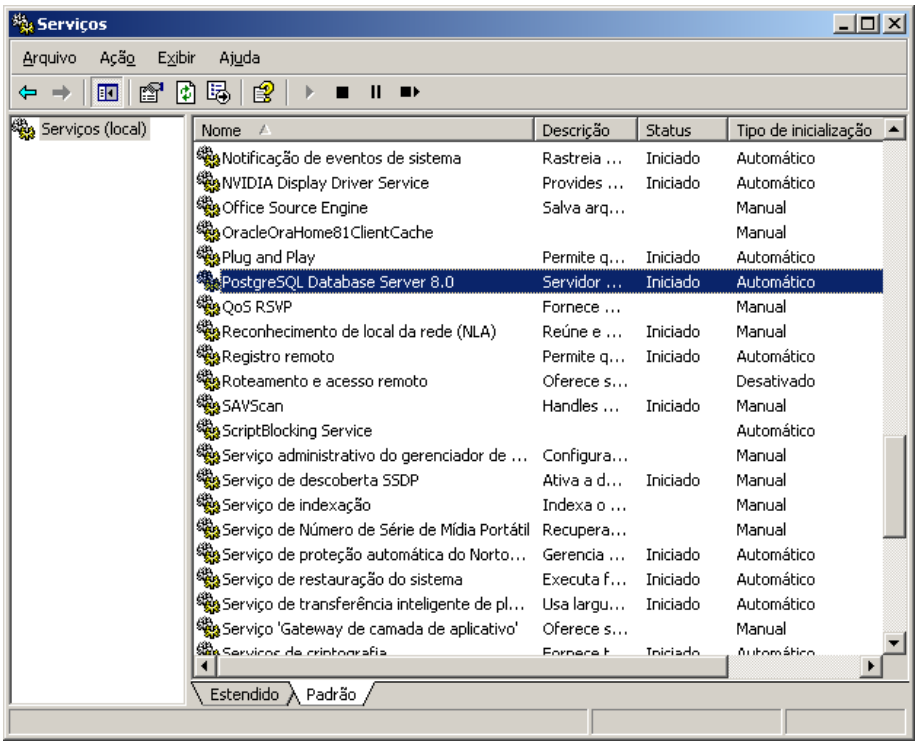

Figura 2.17: Configuração - lista de serviços do computador

Na figura 2.19 é ilustrada a opção para iniciar e a opção para finalizar a execução do servidor de banco de dados PostgreSQL.

2.4 Configurando as permissões do servidor e criando o banco de dados do MIP

O último passo necessário para concluir a instalação do PostgreSQL é configurar as permissões do servidor e criar uma nova base de dados para o sistema MIP. Antes de começar esta etapa, inicie o servidor, conforme descrito na seção anterior.

Agora é necessário executar o programa de administração do PostgreSQL. O programa de

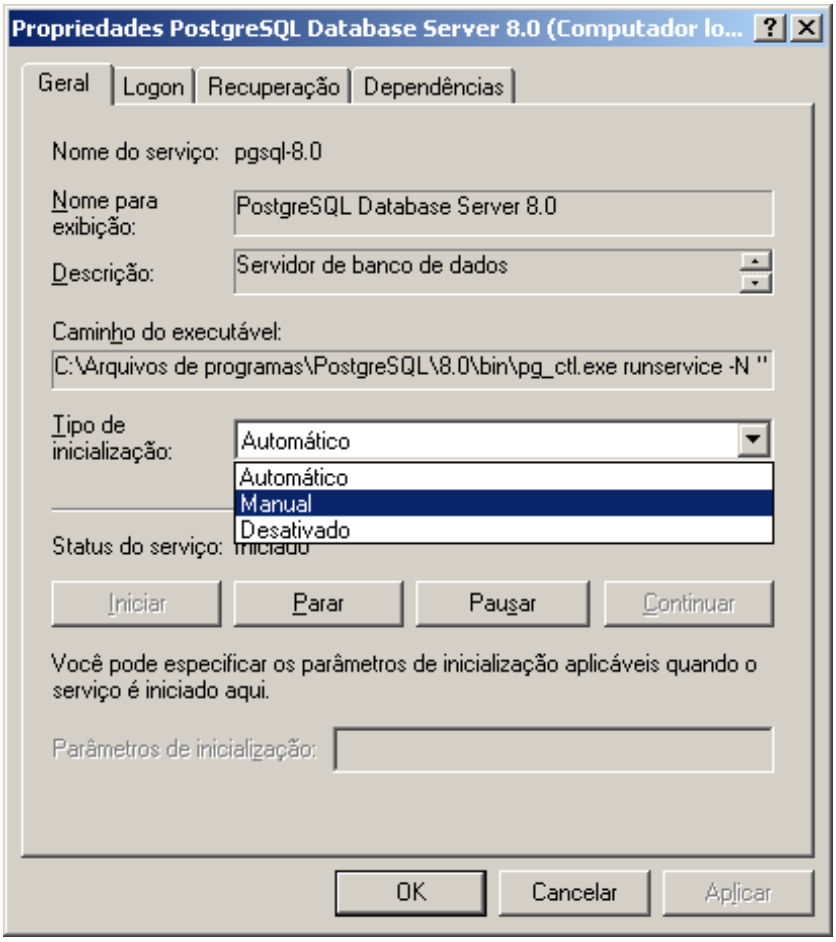

Figura 2.18: Configuração - alteração do tipo de inicialização do serviço

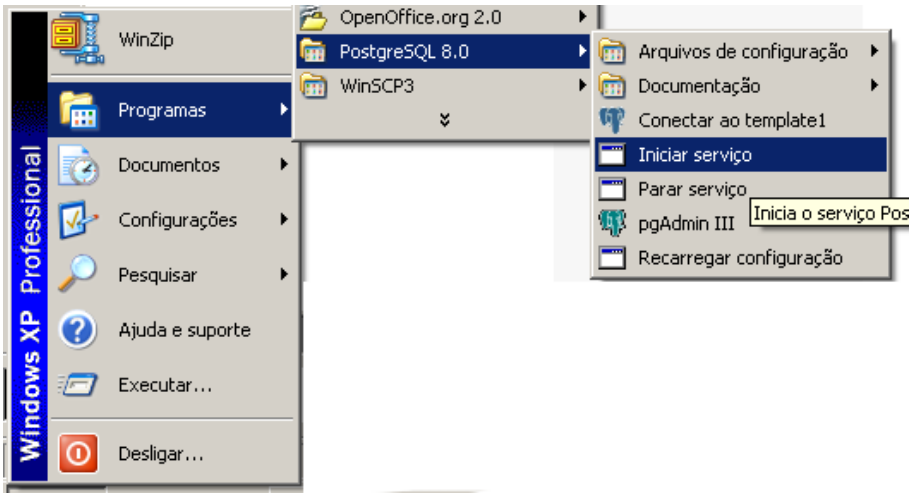

Figura 2.19: Menu de opções no qual é possível iniciar e parar o servidor de banco de dados

administração do PostgreSQL pode ser executado através da opção “pgAdmin III”, disponível no menu “Iniciar” do Windows e ilustrada na figura 2.20.

A tela principal do “pgAdmin III” é ilustrada na figura 2.21. Por padrão, existe um banco de dados já configurado na listagem à esquerda da tela. Trata-se do banco de dados que a pouco foi instalado. Para acessá-lo, clique duas vezes no nome “PostgreSQL Database Server 8.0” na listagem do lado esquerdo da tela da figura 2.21.

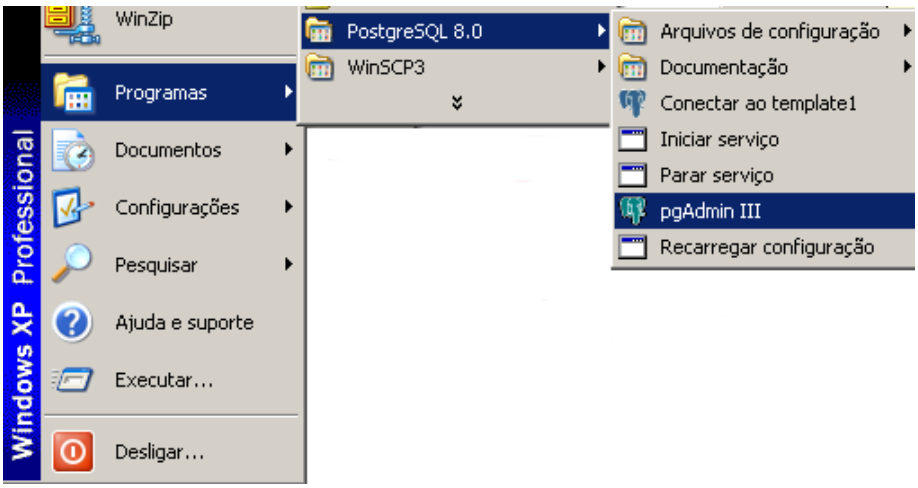

Figura 2.20: Configuração - executando o “pgAdmin III”

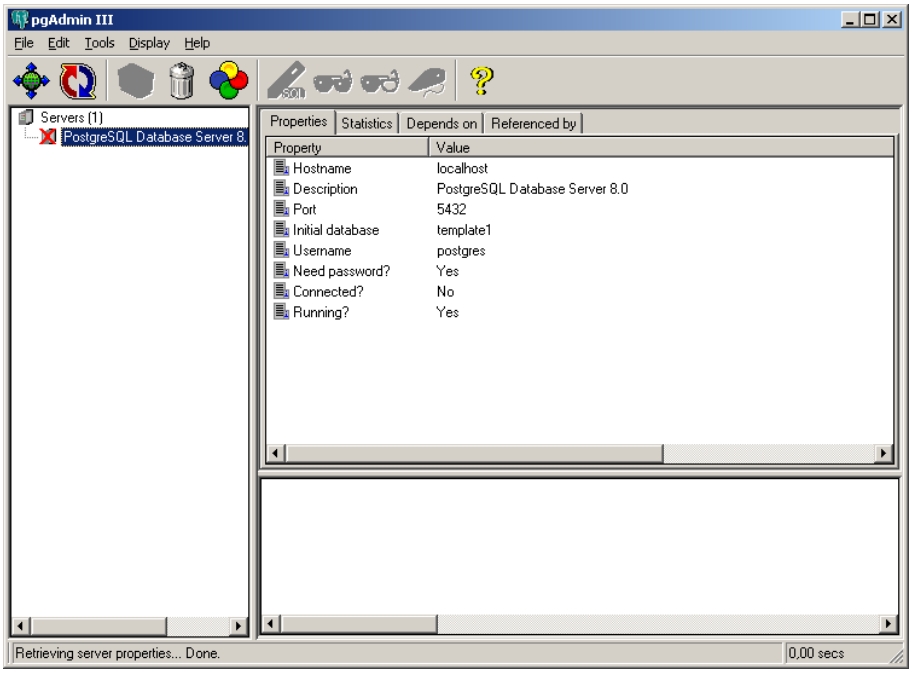

Figura 2.21: Configuração - tela inicial do “pgAdmin III”

O sistema solicitará a senha do usuário administrador (usuário postgres), conforme ilustrado na figura 2.22. Digite a senha criada na tela exibida na figura 2.8. Após digitá-la, clique em “OK”.

Figura 2.22: Configuração - digitando a senha de acesso para administrar o servidor

A tela inicial do “pgAdmin III” será novamente exibida, porém com mais itens na listagem à esquerda (figura 2.23). O primeiro passo é criar um usuário para utilização do MIP. Na listagem à esquerda da tela do “pgAdmin III”, clique com o botão direito do mouse sobre a opção “Users”. Aparecerá um menu, conforme ilustrado na figura 2.23. Selecione a opção “New User”.

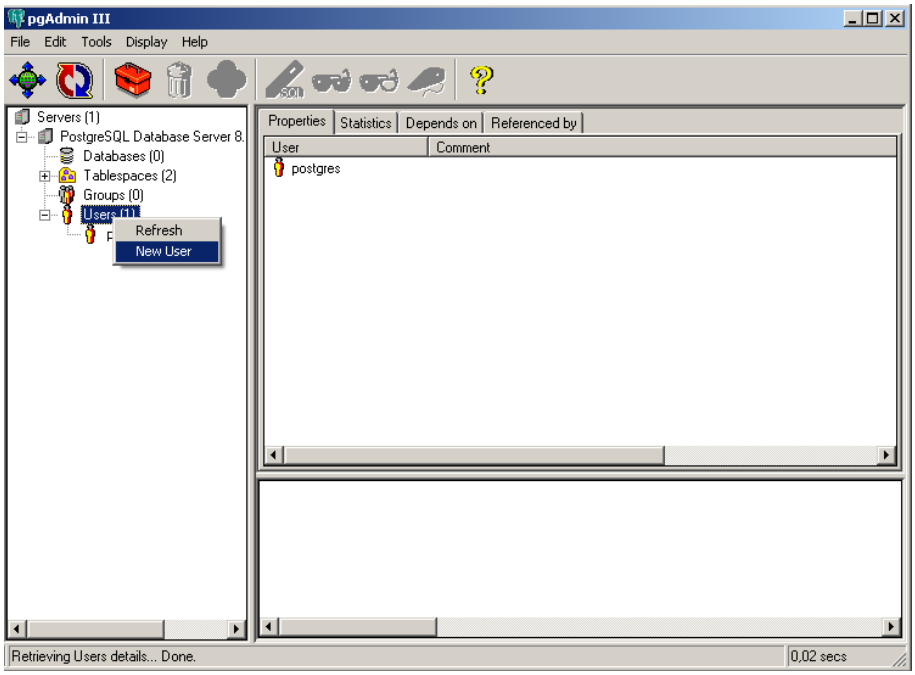

Figura 2.23: Configuração - “pgAdmin III” - criando um novo usuário

Após selecionar a opção “New User” será exibida a tela para criação do novo usuário. Esta tela é ilustrada na figura 2.24. Na caixa “Username” digite “mip”. Na caixa “Password” digite “conversao”.

Caso deseje alterar o nome do usuário ou a senha padrão do sistema MIP, edite o arquivo “JDBC.propriedades” que se encontra no diretório “conf” da instalação do sistema. Neste arquivo existirão duas linhas com os seguintes comandos:

```
jdbc.usuario=mip

jdbc.senha=conversao
```

O nome do usuário e a senha digitados na tela da figura 2.24 devem ser os mesmos que os digitados nestas duas linhas do arquivo “JDBC.propriedades”. Caso você escolha outro nome de usuário ou outra senha de acesso, basta alterar as duas linhas do arquivo de propriedades do MIP e salvá-lo com as alterações.

Agora é necessário criar o banco de dados do sistema MIP. Na tela principal do “pgAdmin III” clique com o botão direito do mouse sobre a opção “Databases”, conforme ilustrado na figura 2.25. Selecione a opção “New Database”.

Será exibida a tela da figura 2.26. Nesta tela devem ser preenchidos os campos “Name” e “Owner”. O campo “Owner” deve conter o nome do usuário criado no passo anterior. O campo “Name” deve conter o nome do banco de dados do sistema MIP.

Por padrão, o nome do banco de dados do MIP é “conversaomip”. Caso seja necessário alterar este nome, é preciso alterar também o arquivo “JDBC.propriedades”. No arquivo

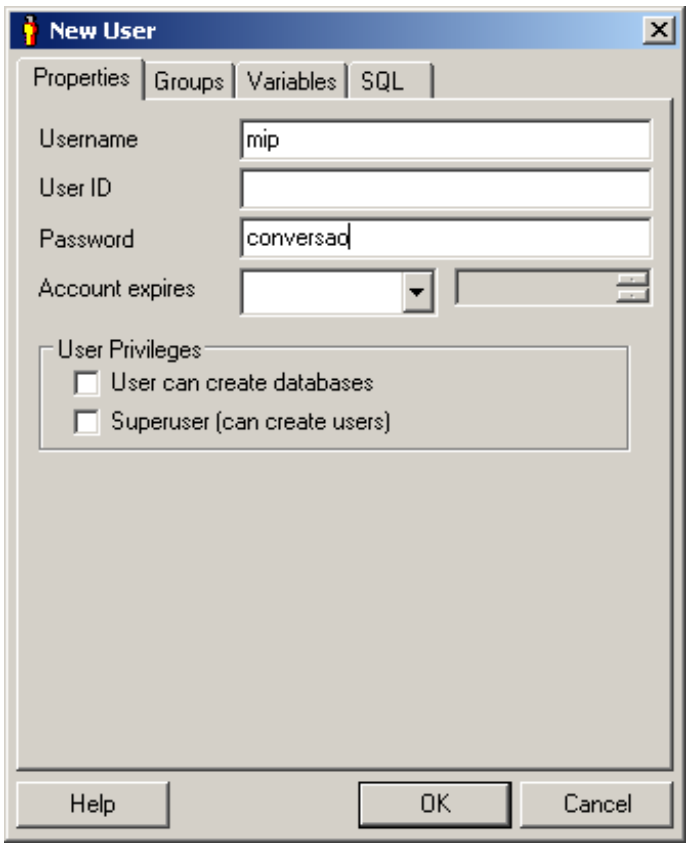

Figura 2.24: Configuração - tela de criação do usuário do MIP no servidor

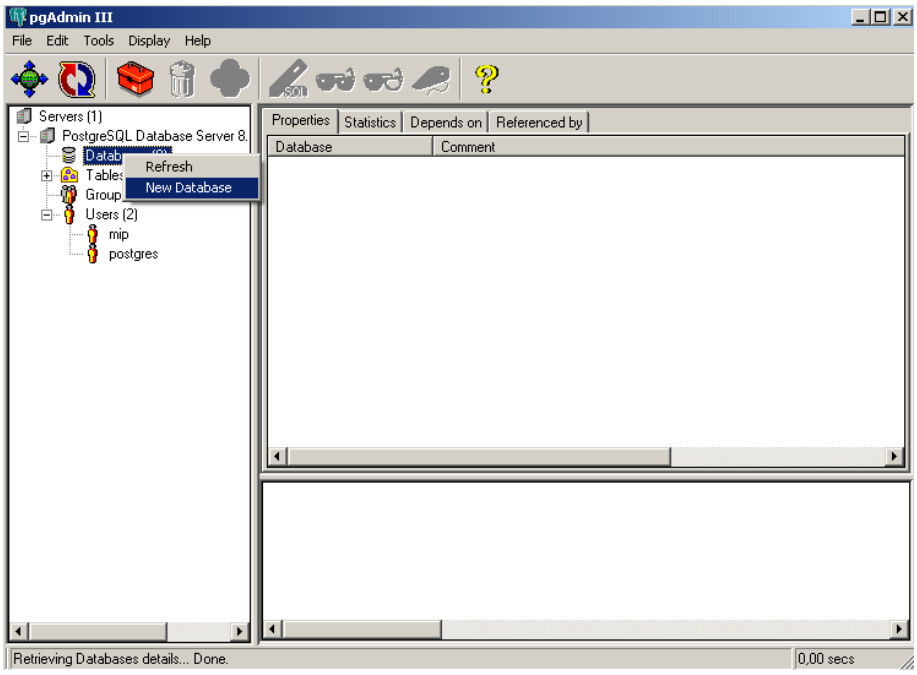

Figura 2.25: Criando um novo banco de dados

"JDBC.propriedades", altere a linha que contém o comando abaixo para que o final da linha contenha o mesmo nome do banco de dados, digitado na tela da figura 2.26.

jdbc.url=jdbc:postgresql://localhost/conversaomip

O sistema agora está pronto para ser utilizado. Feche o "pgAdmin III" e execute o MIP,

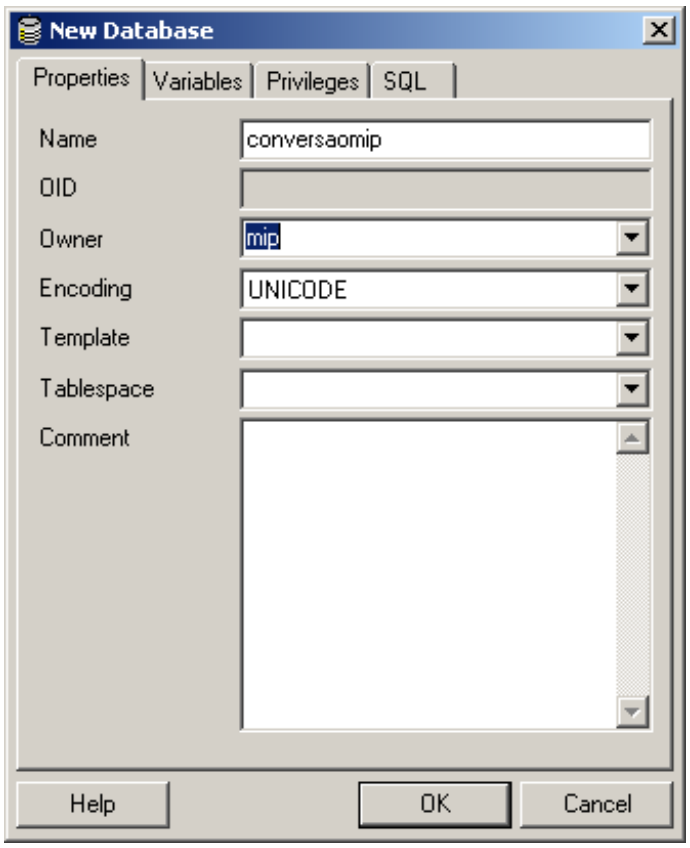

Figura 2.26: Preenchendo as informações do novo banco de dados

conforme descrito na próxima seção. Lembre-se de sempre iniciar o PostgreSQL antes de usar o sistema, e finalizar o PostgreSQL quando não estiver mais executando o MIP. Isso melhorará o desempenho de seu computador.

2.5 Executando o sistema

Dentro da pasta do sistema MIP estão os executáveis para cada módulo do sistema, tanto para o sistema operacional Windows, quanto para o sistema operacional Linux.

Para executar o sistema em plataforma Windows, clique em:

treinamento.bat - para iniciar o Módulo de Aprendizagem

consulta.bat - para iniciar o Módulo de Consultas

Para executar o sistema em plataforma Linux, execute o arquivo:

treinamento.sh - para iniciar o Módulo de Aprendizagem

consulta.sh - para iniciar o Módulo de Consultas

Lembre-se que, caso vá executar uma conversão de dados do formato de saída do Datawarehouse, é necessário iniciar também o PostgreSQL.

Capítulo 3

Módulo de Aprendizagem

O Módulo de Aprendizagem é a parte do sistema responsável por realizar o treinamento de um classificador a partir da base de dados selecionada pelo usuário. Um classificador é um componente do sistema capaz de realizar a classificação de clientes como candidatos à inspeção ou não. Após o classificador ter sido gerado, o Módulo de Aprendizagem permite exportá-lo para utilização no Módulo de Consultas.

Este módulo será apresentado basicamente em três seções: a primeira exibirá as interfaces e descreverá como elas são utilizadas; a segunda apresentará as principais funcionalidades disponíveis nesta parte do sistema; a terceira descreverá cada técnica de classificação e validação disponível e os seus parâmetros configuráveis.

3.1 Conhecendo as interfaces do Módulo de Aprendizagem

Nesta seção serão apresentadas as interfaces do Módulo de Aprendizagem e será descrito como cada uma delas é utilizada.

3.1.1 Janela inicial

A interface inicial deste módulo é ilutrada na figura 3.1. A partir desta interface tem-se acesso às funcionalidades do sistema.

3.1.2 Menu Arquivo

Uma configuração de treinamento é o conjunto formado pela técnica escolhida pelo usuário (com suas suas opções), o estimador de erro selecionado (também com suas suas opções) e a base de treinamento escolhida. O conjunto formado por várias combinações destes três elementos é exibido na fila de execução de tarefas da tela principal do Módulo de Aprendizagem. Este conjunto pode ser salvo e restaurado para futuras sessões do sistema através do menu “Arquivo”, que é ilustrado na figura 3.2.

O menu “Arquivo” possui 5 opções, que realizam as seguintes tarefas:

- “Nova configuração” - permite criar uma nova configuração.
- “Abrir configuração” - permite abrir uma configuração existente.
- “Salvar configuração” - permite salvar a configuração atual.
- “Salvar como” - permite salvar a configuração atual com outro nome desejado.

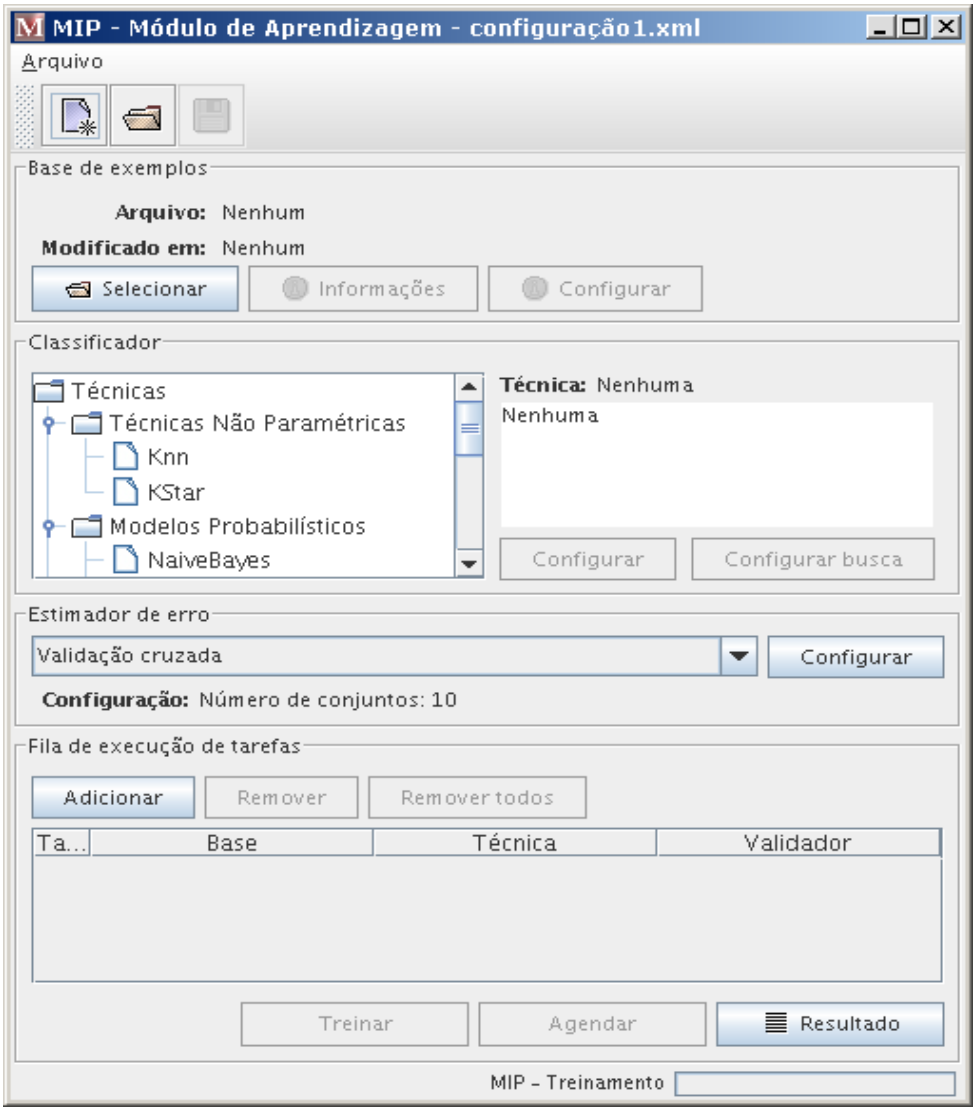

Figura 3.1: Tela inicial do Módulo de Aprendizagem

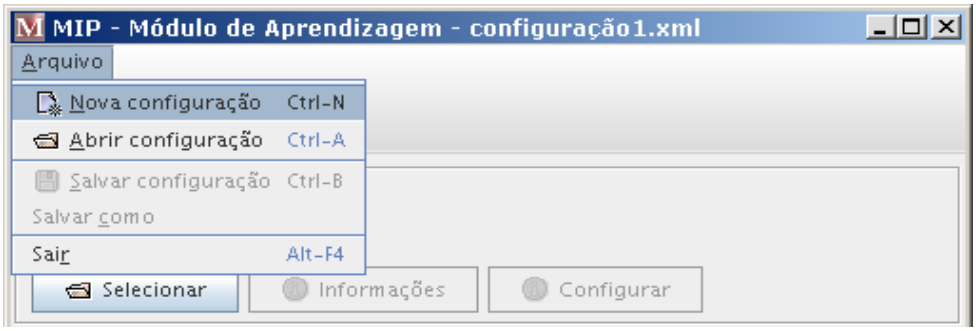

Figura 3.2: Menu “Arquivo” no qual é possível manipular configurações

- “Sair” - permite ao usuário sair do Módulo de Aprendizagem.

Algumas destas opções do menu “Arquivo” estão também disponíveis na barra de botões do topo da tela. Esta barra de botões é mostrada na figura 3.3.

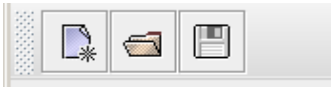

Figura 3.3: Botões “Novo”, “Abrir” e “Salvar”

3.1.3 Seleção da base de dados

No início da parte central da tela principal do sistema encontra-se o painel de manipulação da base de dados de treinamento. Este painel possui três botões que permitem selecionar uma base de dados, obter informações e alterar configurações da base selecionada. A figura 3.4 ilustra o painel de manipulação da base.

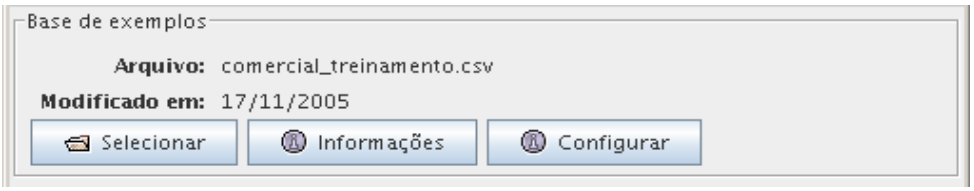

Figura 3.4: Painel de seleção de base de exemplos

O funcionamento desta parte do sistema é descrito em detalhes na capítulo 5.

3.1.4 Seleção e configuração do classificador

No painel “Classificador” a árvore de técnicas permite que seja selecionada a técnica de classificação a ser utilizada no treinamento com a base de dados. Na caixa à direita será exibida a configuração padrão da técnica selecionada. O botão “Configurar” permite que esta configuração seja alterada.

Na figura 3.5 a técnica BayesNetB foi selecionada na árvore de técnicas e o painel apresenta as configurações da técnica, por exemplo: iniciar como Naive Bayes (sim/não), número máximo de pais a serem usados no classificador, o tipo de medição (MDL/AIC) da avaliação do resultado do classificador e o parâmetro Alpha de ajuste do classificador.

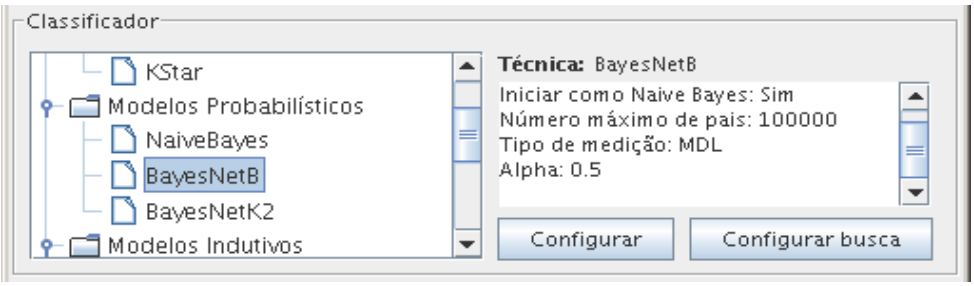

Figura 3.5: Painel de seleção do classificador

O botão “Configurar busca” permite configurar uma busca pelos melhores parâmetros da técnica selecionada, ou seja, uma pesquisa por uma configuração que gere boas estatísticas para o classificador.

Esta opção só será habilitada se a técnica possuir parâmetros configuráveis. O procedimento a ser executado para realização de uma busca por parâmetros é explicado, juntamente com as

possíveis configurações da busca, na seção 3.2.7.

3.1.5 Seleção e configuração do estimador de erro

No painel “Estimador de erro”, ilustrado na figura 3.6, é exibida uma caixa para seleção do tipo de validação do classificador, ou seja, qual a estratégia de treinamento e de teste será usada na criação do classificador. Os diferentes tipos de validadores são explicados na seção 3.2.6.

O botão “Configurar” do painel permite ajustar as opções do estimador selecionado. No caso ilustrado na figura 3.6, em que foi selecionado o estimador “Validação Cruzada”, este botão permite escolher o número de conjuntos da validação.

Figura 3.6: Painel de escolha do estimador de erro

3.1.6 Fila de execução de tarefas

A fila de execução de tarefas listará todas as tarefas que já estão devidamente configuradas e que estão prontas para serem executadas. A figura 3.7 mostra a fila de execução de tarefas e os botões de controle para manipulação da fila.

Figura 3.7: Fila de execução de tarefas

O botão “Adicionar” permite adicionar uma nova tarefa à fila. Isto só será possível se já tiverem sido selecionadas uma base, uma técnica e um estimador de erro.

Caso o botão “Configurar busca” tenha sido o último acionado após a seleção de uma técnica, a nova tarefa será criada como uma busca por melhores parâmetros da técnica em questão ao invés de uma tarefa de treinamento normal.

O botão “Remover” permite remover uma tarefa da fila de execução e o botão “Remover todos” retira todas as tarefas da fila.

Além destes botões, o painel da fila de execução ainda disponibiliza os botões “Treinar”, “Agendar” e “Resultado”.

O botão “Treinar” coloca em execução as tarefas que estão na fila e mostra a janela do “Gerenciador de classificadores” ou a janela do “Gerenciador de buscas” quando todas as tarefas estive-

rem concluídas. Durante a execução das tarefas, o botão “Treinar” se transforma em “Cancelar”. O acionamento do botão “Cancelar” interrompe e execução atual imediatamente. Os resultados obtidos até então não serão perdidos.

O botão “Agendar” permite ao usuário salvar a lista de tarefas para execução posterior. A lista gerada pode ser agendada para execução em um servidor, por exemplo. O funcionamento desta opção é detalhado na seção 3.2.3.

O botão “Resultado” exibe o “Gerenciador de classificadores” ou o “Gerenciador de buscas”, conforme ilustrado na figura 3.8. O funcionamento destes dois gerenciadores será detalhado nas duas próximas seções.

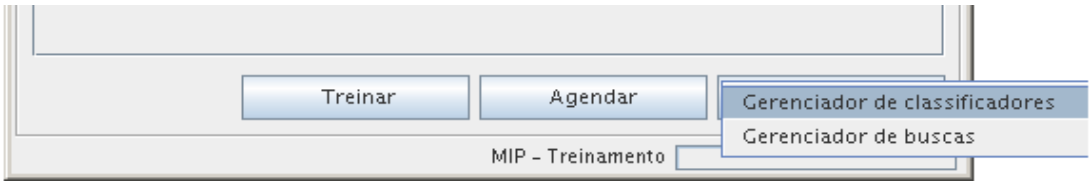

Figura 3.8: Escolha entre o gerenciador de classificadores e o gerenciador de buscas na tela principal do Módulo de Aprendizagem

A barra de estado, na parte inferior da fila de execução, exibe mensagens que informam o estado atual da execução do treinamento.

3.1.7 Gerenciador de Classificadores

Esta parte do sistema exibe os classificadores gerados e as respectivas estatísticas de desempenho. Conforme ilustrado na figura 3.9, esta interface possui quatro botões de controle.

O botão “Exportar” permite que um classificador selecionado seja exportado para uso no Módulo de Consultas. O botão “Exportar todos” realiza a mesma tarefa, porém exportando toda a lista de classificadores disponível. O botão “Remover” permite ao usuário eliminar do gerenciador o classificador atualmente selecionado. O botão “Fechar” fecha o Gerenciador de Classificadores, retornando à janela principal do Módulo de Aprendizagem.

Na lista de classificadores será exibido o nome da técnica, a categoria do algoritmo e a base de dados do treinamento de cada classificador já construído. Os classificadores aparecem na lista de acordo com a ordem em que eles foram gerados, do mais antigo para o mais recente.

O painel “Informações” exibe informações gerais e as estatísticas que demonstram o desempenho do classificador. Dentre as informações gerais do classificador, são exibidos o nome da técnica, a categoria da técnica, a data e a hora em que foi realizado o treinamento, e as opções (configurações) da técnica utilizada no treinamento.

Nas informações sobre a base de exemplos é exibido o nome, a descrição, o tipo, o total de exemplos e o total de atributos da base.

Na parte de estatísticas de desempenho é exibido o estimador de erro utilizado, o número de instâncias do treinamento, o número de classificações corretas, o número de classificações incorretas, o percentual de acerto, a especificidade, a confiabilidade negativa, a média harmônica entre a especificidade e a confiabilidade e a matriz de confusão do classificador.

A especificidade do classificador é o número de fraudadores classificados como fraudadores sobre o total de fraudadores existentes na base. A confiabilidade negativa é o número de fraudadores classificados como fraudadores sobre o total de exemplos classificados como fraudadores.

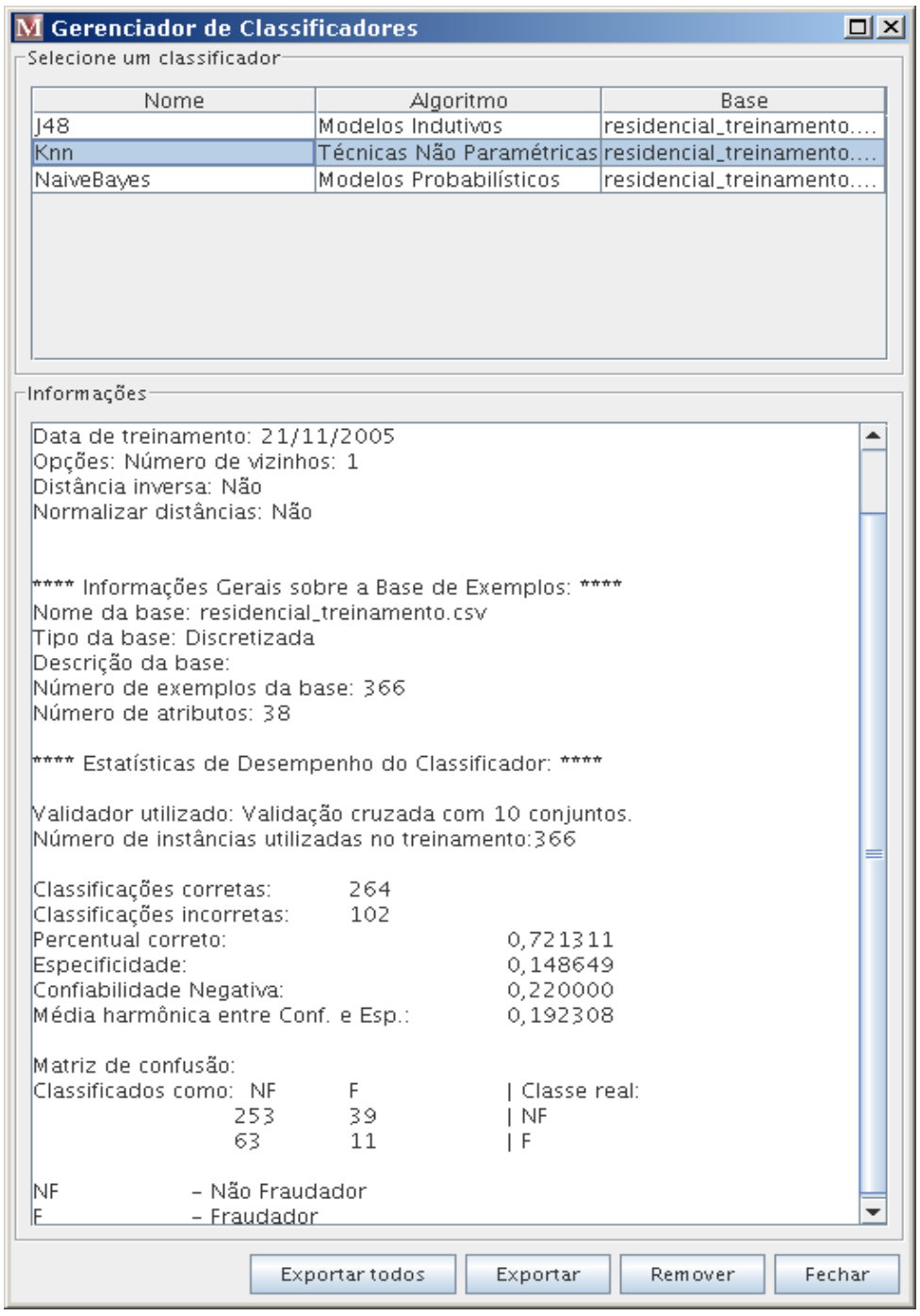

Figura 3.9: Gerenciador de classificadores

A matriz de confusão mostra o total de casos da base de acordo com as classes reais e as classes preditas (fraudadores e não-fraudadores), oferecendo uma maneira eficiente de verificar o desempenho do classificador.

3.1.8 Gerenciador de Buscas

Esta interface exibe as buscas realizadas e as estatísticas de desempenho dos classificadores validados durante a busca. No painel de seleção de busca são exibidas, para cada busca, as seguintes informações: o nome da técnica utilizada, a classe de algoritmo à qual a técnica pertence e a base utilizada.

Um exemplo é mostrado na figura 3.10. O painel de resultados da busca mostra, para a busca

que estiver selecionada, uma lista das configurações avaliadas. Para cada configuração são exibidos: o número da iteração em que a configuração foi avaliada, as opções da configuração e as medidas de desempenho que a configuração obteve (taxa de acerto, especificidade, confiabilidade negativa e o valor da métrica utilizada).

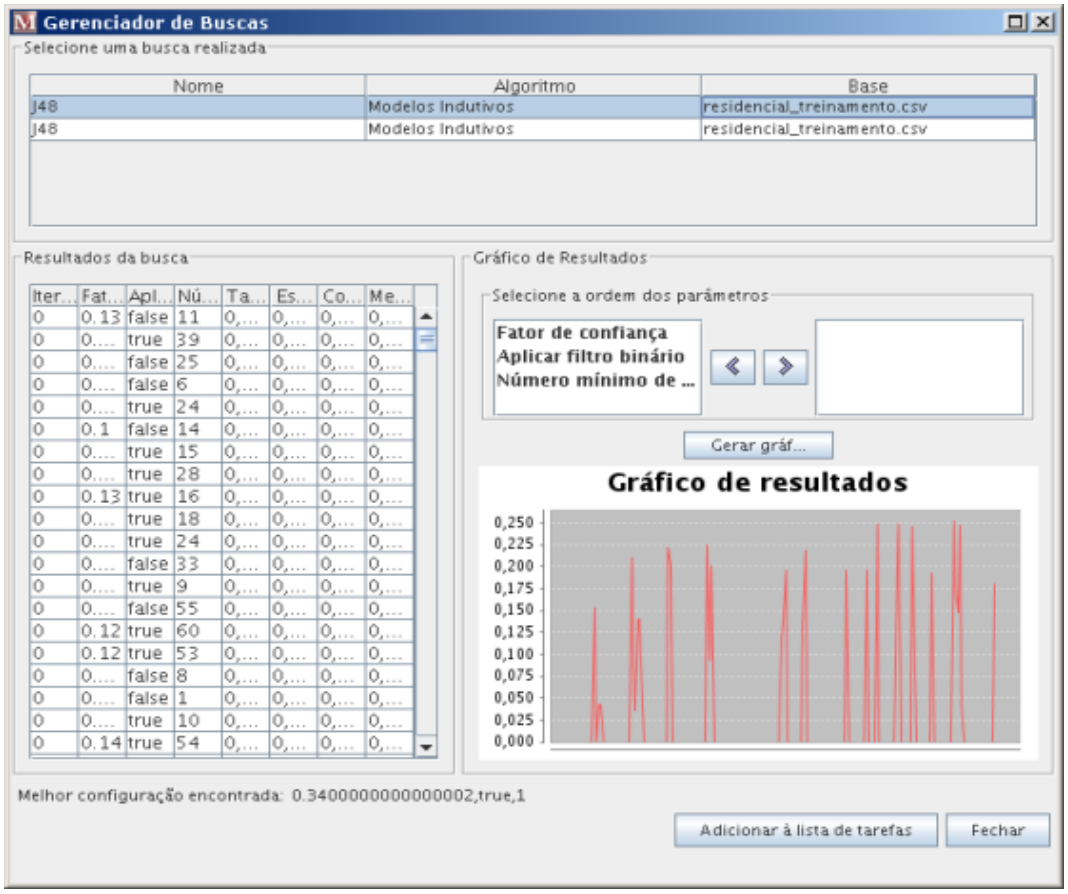

Figura 3.10: Gerenciador de buscas

No painel do gráfico de resultados é gerado um gráfico do tipo configuração X desempenho. No eixo X aparecem as configurações avaliadas e no eixo Y o valor obtido para a métrica em cada configuração. As configurações no eixo X são ordenadas de acordo com a ordem escolhida no painel de seleção de ordem dos parâmetros. Assim, elas serão ordenadas primeiro pelos valores do primeiro parâmetro e, em caso de empate, pelos valores do segundo parâmetro e assim por diante.

Abaixo do painel de resultados é mostrada a melhor configuração avaliada durante o processo de busca. O botão “Adicionar à lista de tarefas” permite que a melhor configuração encontrada ou a configuração atualmente selecionada na lista de resultados seja adicionada à lista de tarefas da janela principal do Módulo de Aprendizagem. Caso o usuário deseje exportar algumas destas configurações para o Módulo de Consultas, será necessário adicionar as configurações desejadas à lista de tarefas e clicar novamente no botão “Treinar” na janela principal do Módulo de Aprendizagem.

O botão “Fechar” fecha o Gerenciador de Buscas, retornando à janela principal do Módulo de Aprendizagem.

3.2 Funcionalidades do Módulo de Aprendizagem

Nesta seção são descritos os passos para execução das principais funções do Módulo de Aprendizagem.

3.2.1 Criando uma tarefa de treinamento

Para inserir uma nova tarefa de treinamento à lista de tarefas é necessário executar os seguintes passos:

1. Selecionar uma base de exemplos - usando o painel de seleção de base de dados, mostrado na figura 3.4, o usuário deve selecionar uma base de dados para a nova tarefa de treinamento. O usuário pode optar por selecionar uma base de dados em formato CSV ou abrir a tela de conversão de dados de saída do Datawarehouse da ESCELSA. O funcionamento desta parte do sistema é descrito em detalhes na capítulo 5.

   Além disso, o usuário possivelmente necessitará configurar as propriedades da base selecionada atreves do botão “Configurar”. Os detalhes de como realizar esta configuração também encontram-se no capítulo 5.

   Um detalhe importante é que cada tarefa da lista de execução pode ser criada com uma base diferente. Neste caso, basta configurar corretamente cada base após selecioná-la (ver capítulo 5).

2. Selecionar uma técnica e configurar suas opções - para selecionar uma técnica o usuário deve ir ao painel “Classificador” e escolher a técnica desejada na árvore de técnicas disponíveis. O sistema exibirá no lado direito da tela o nome da técnica e suas respectivas configurações. O painel de seleção de classificadores é mostrado na figura 3.11.

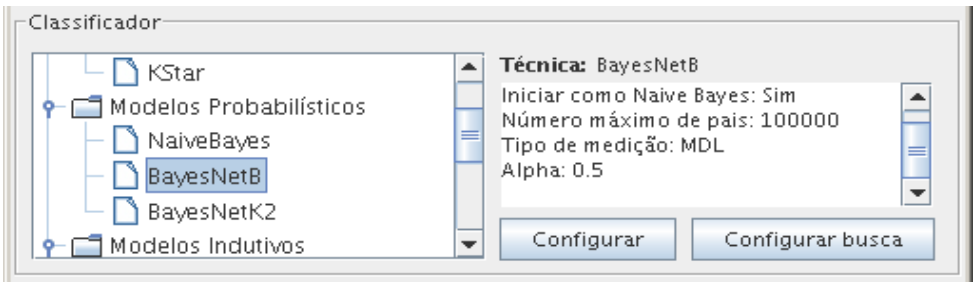

Figura 3.11: Painel de seleção do classificador

   Caso seja necessário alterar as configurações da técnica selecionada, basta clicar no botão “Configurar”. Como exemplo, para a técnica BayesNetB aparecerá a tela de opções mostrada na figura 3.12.

   Caso o objetivo seja realizar a busca por melhores parâmetros da técnica, deve-se clicar no botão “Configurar busca” e alterar, se desejado, as configurações da busca. A tela com os parâmetros da busca e a explicação do significado de cada parâmetro serão mostrados na seção 3.2.7.

3. Selecionar um estimador de erro - essa funcionalidade permite ao usuário escolher a técnica que irá validar o desempenho do classificador a ser gerado. O usuário poderá escolher um

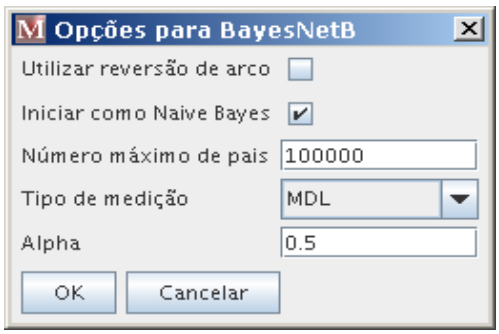

Figura 3.12: Janela de opções da técnica BayesNetB

dos validadores disponíveis no painel “Estimador de erro”, conforme mostrado na figura 3.13.

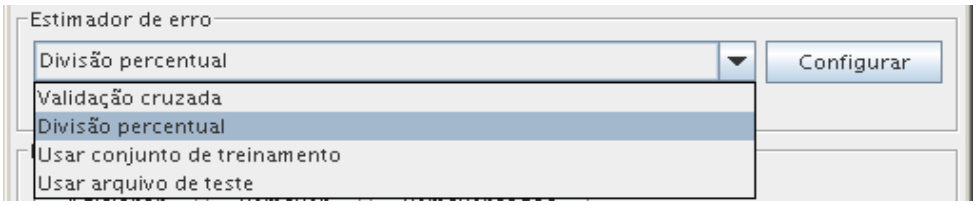

Figura 3.13: Painel para a escolha do estimador de erro

Após selecionar o validador, é possível alterar a sua configuração acionando o botão “Configurar”. Como exemplo, na figura 3.14 é mostrada a tela de propriedades do validador “Divisão Percentual”.

Figura 3.14: Janela de opções para o estimador de erro “Divisão percentual”

4. Adicionar a nova tarefa à lista de execução - após selecionar uma base de dados, uma técnica de classificação e um estimador de erro, o usuário pode adicionar a nova tarefa à fila de execução acionando o botão “Adicionar”. É importante ressaltar que não há restrição sobre o limite de tarefas na fila de execução e nem sobre o tipo de tarefas que forma a fila: para uma mesma execução podem ser incluídas tarefas tanto de treinamento quanto de busca de parâmetros.

3.2.2 Executando um treinamento

Após criar uma lista de tarefas, o usuário pode iniciar o treinamento acionando o botão “Treinar”. Neste momento, o botão será substituído por um botão “Cancelar”, conforme mostrado na figura 3.15.

Se o botão “Cancelar” for acionado, a execução da tarefa atual será interrompida e as tarefas posteriores não serão executadas. Caso a tarefa atual seja uma busca por melhores parâmetros,

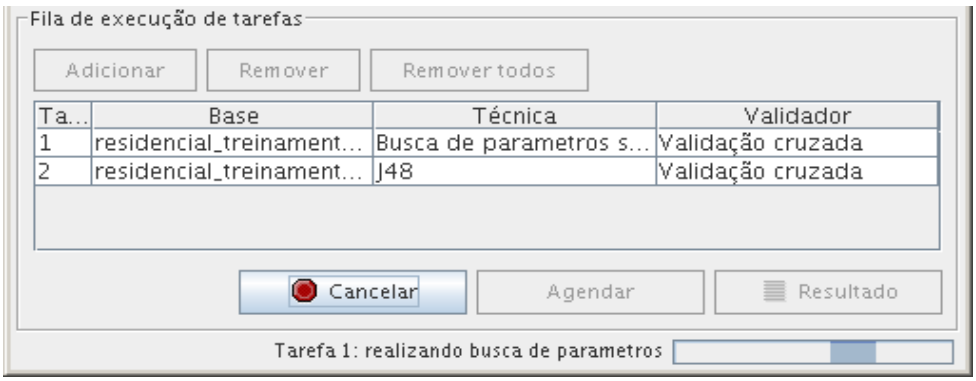

Figura 3.15: Fila com tarefas em execução

será exibida uma lista com as configurações testadas até então, conforme ilustrado na figura 3.16. Os parâmetros da lista podem ser aproveitados para configuração de novos classificadores manualmente, conforme mostrado na seção 3.2.6. Caso alguma das tarefas já executada seja de treinamento, o classificador gerado estará disponível na interface de Gerência de Classificadores.

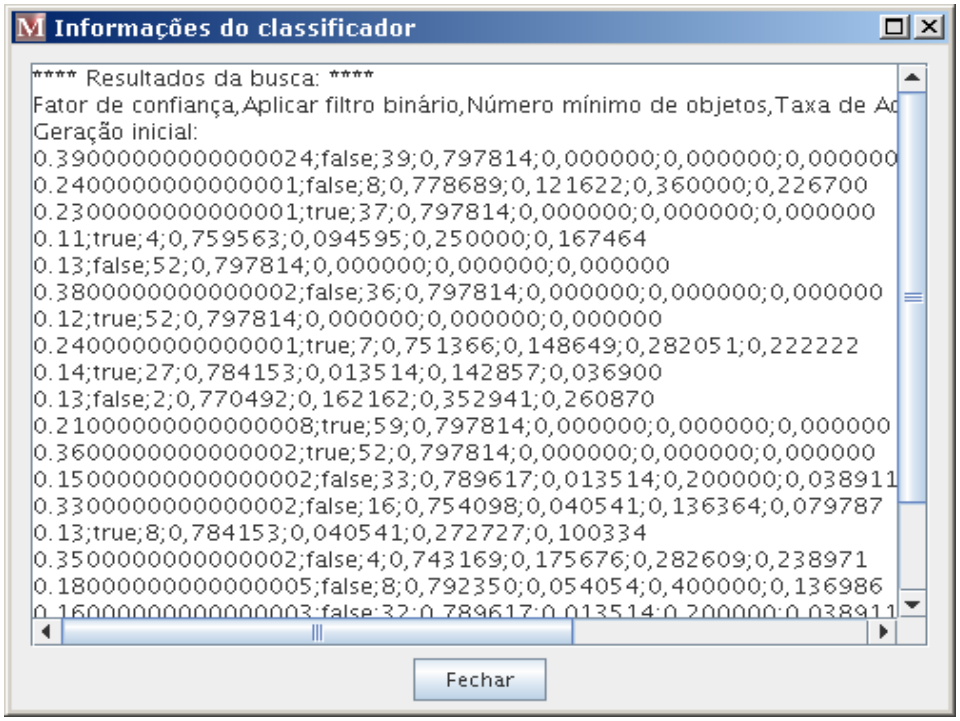

Figura 3.16: Lista de configurações testadas em um busca interrompida

3.2.3 Agendando a execução do treinamento

Essa funcionalidade permite que as configurações que estão na fila de execução de tarefas possam ser exportadas para execução posterior. Para agendar a execução de uma lista de tarefas é necessário clicar no botão “Agendar” da interface principal. Será exibida a tela ilustrada na figura 3.17, para seleção do arquivo no qual será salva a fila de execução.

Após digitar o nome de um arquivo *XML*, o sistema exibirá a mensagem ilustrada na figura 3.18, que informa ao usuário o comando para iniciar a execução da fila exportada. Neste momento, se for de interesse do usuário, o sistema pode ser fechado que a configuração não será mais perdida.

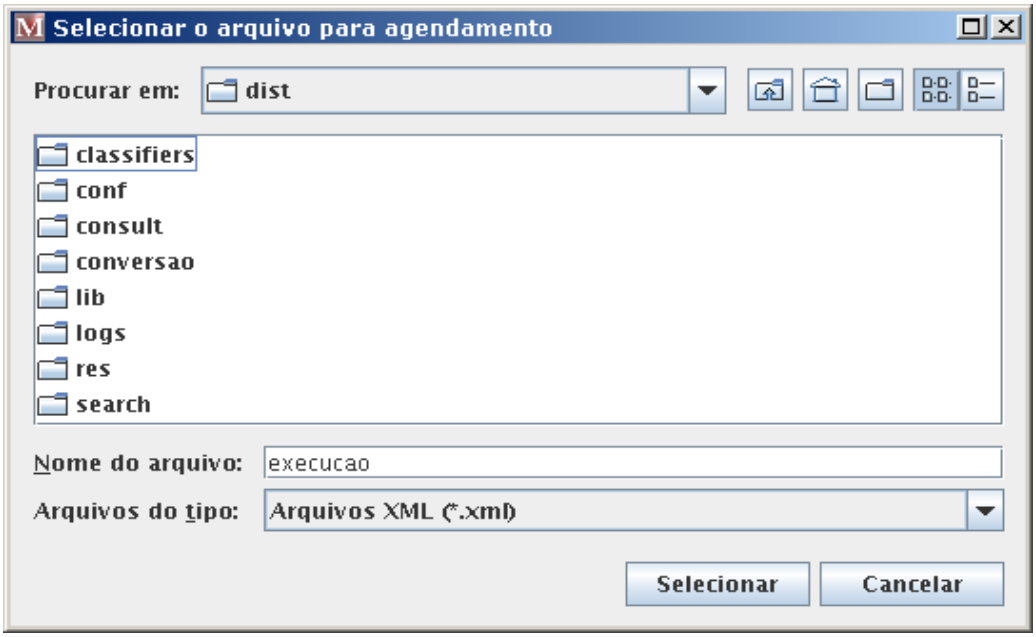

Figura 3.17: Seleção de arquivo que conterá a lista de tarefas

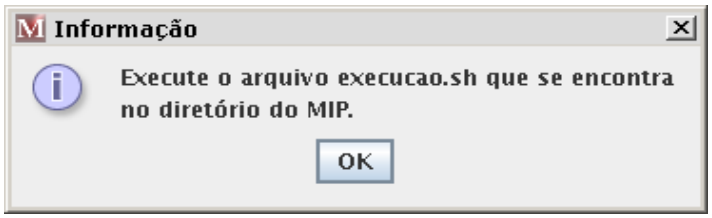

Figura 3.18: Mensagem mostrado qual arquivo deve ser executado para iniciar o treinamento

Caso seja de interesse do usuário executar a fila salva em algum outro computador, é necessário copiar para esta outra máquina os arquivos das bases de dados usadas na criação da fila de execução, o arquivo XML salvo na interface da figura 3.17 e o arquivo cujo nome foi apresentado na interface da figura 3.18 (neste exemplo, *execucao.sh*).

A máquina que executará a fila salva deve possuir o MIP instalado (ver seção 2) e os arquivos copiados deverão estar em uma pasta com o mesmo nome e com o mesmo caminho que eles estavam na máquina onde a lista de tarefas foi gerada. Por exemplo, se na máquina onde a lista de tarefas foi gerada as bases estavam no diretório *c:/dados*, na máquina de execução, as bases deverão estar exatamente neste mesmo diretório.

Para executar uma lista salva, basta utilizar o arquivo cujo nome foi mostrado na interface da figura 3.18 (neste exemplo, *execucao.sh*). Não é necessário estar com o sistema aberto. A fila será executada em uma janela do sistema operacional, sem que a interface do MIP seja carregada.

3.2.4 Verificando resultados e exportando um classificador

Durante a geração de um classificador, um processo de validação é executado para avaliação do desempenho esperado deste classificador. O sistema disponibiliza ao usuário as estatísticas obtidas nesta validação na interface de gerência de classificadores. Esta interface é acessada através do botão “Resultados”, conforme ilustrado na figura 3.8.

Na tela de gerência de classificadores estão disponíveis várias estatísticas, descritas na seção

3.1.7. Para acessar estas estatísticas, basta selecionar uma das linhas da tabela de classificadores, conforme ilustrado na figura 3.19.

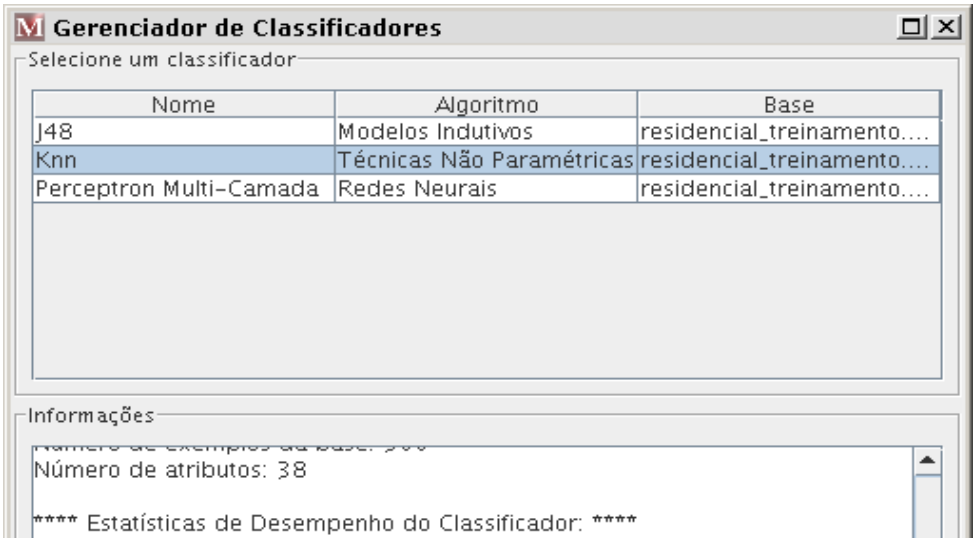

Figura 3.19: Lista de classificadores disponíveis

Os classificadores com melhor desempenho podem ser selecionados pelo usuário para serem exportados para uso no Módulo de Consultas. As informações do classificador serão salvas em um arquivo *ZIP* que conterá dois arquivos: um com extensão *Model* e outro no formato *XML*.

Para exportar um classificador basta selecioná-lo e clicar no botão "Exportar" na parte inferior da tela de gerência de classificadores, conforme mostrado na figura 3.20. Será exibida uma janela para seleção do arquivo no qual o classificador selecionado será salvo.

Figura 3.20: Botões para manipulação dos classificadores

Está disponível também a opção para exportação de todos os classificadores da lista, através do botão "Exportar todos". É ainda possível excluir um classificador da tabela através do botão "Remover", ou ainda clicando com o botão direito do *mouse* na lista de tarefas, conforme mostrado na figura 3.21.

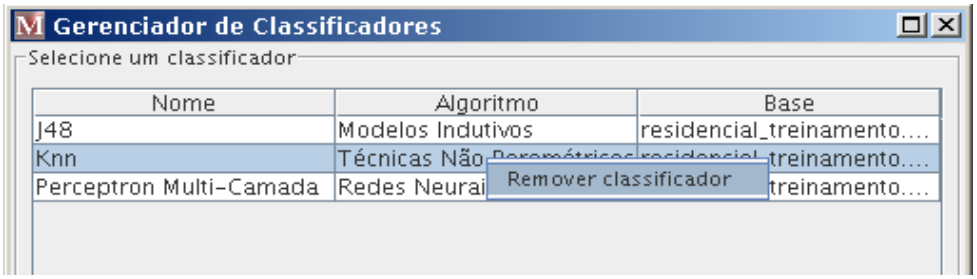

Figura 3.21: Removendo um classificador

A opção "Remover" deve ser usada com cuidado, pois não há outra forma de recuperar um classificador excluído senão executando o treinamento novamente.

3.2.5 Criando, recuperando e salvando configurações

Conforme dito na seção 3.1.2, uma configuração consiste em um conjunto de várias tarefas contendo: uma técnica e suas opções, um estimador de erro e uma bases de dados.

A fila de execução pode ser salva para reutilização posterior. A lista salva será exatamente a mesma salva no caso da utilização da função de agendamento, porém, neste caso, as tarefas salvas deverão ser abertas na interface do MIP e não executadas *fora* do sistema.

Os itens da interface que manipulam configurações estão descritas na seção 3.1.2. Como ilustração, supondo que o sistema tenha sido aberto e que nada tenha sido criado ainda. A fila de execução será exibida como na figura 3.22.

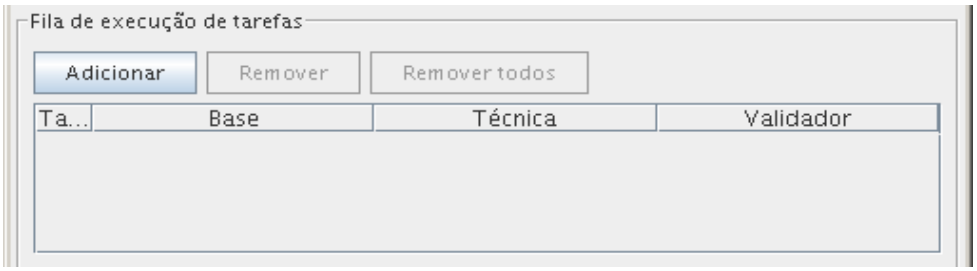

Figura 3.22: Fila de execução vazia

Neste momento, o sistema mostrará que nenhuma configuração ainda foi salva ou aberta através da mensagem na barra de título, mostrada na figura 3.23.

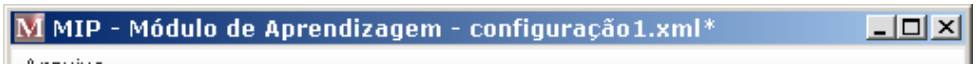

Figura 3.23: Barra de títulos mostra nome da configuração padrão. Nada ainda foi salvo

Após criar uma lista de tarefas o usuário pode salvar a configuração usando os itens do menu “Salvar configuração” ou “Salvar como”, mostrados na seção 3.1.2. Será solicitado o nome do arquivo no qual a lista será salva. A figura 3.24 mostra uma lista de tarefas para ser salva.

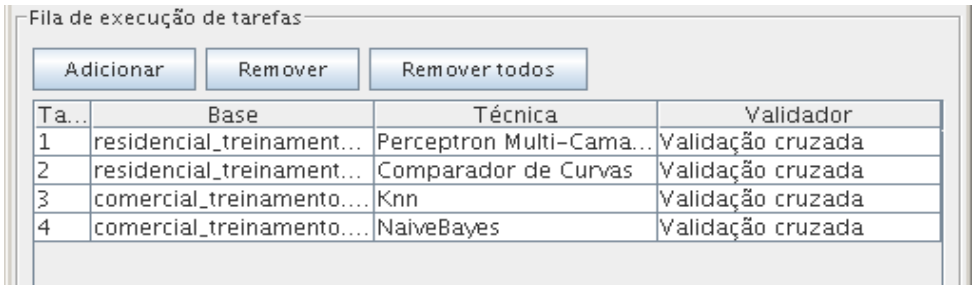

Figura 3.24: Fila de execução com tarefas para serem salvas

Após acionar o menu “Salvar configuração”, o sistema mostrará na barra de título o nome da configuração atualmente carregada, conforme ilustrado na figura 3.25. Notar que neste caso não existe o caractere * em frente ao nome da configuração. A ausência deste caractere indica que nenhuma alteração foi feita na configuração em relação ao que está salvo em disco.

Acionando o item de menu “Nova configuração”, o usuário limpa a lista de tarefas e volta a situação ilustrada nas figuras 3.22 e 3.23.

Figura 3.25: Barra de títulos mostra nome de configuração atualmente carregada

Caso queira recuperar uma configuração já salva, basta utilizar o item de menu “Abrir configuração”. No momento de carregar a configuração, o sistema verifica se a base de dados utilizada na criação das tarefas ainda existe e também se todas as tarefas eram de treinamento com a base de dados atualmente carregada. Se algumas das tarefas não satisfizer estas condições, uma mensagem é apresentada ao usuário perguntando se o sistema deve substituir a base das tarefas salvas. Esta mensagem é ilustrada na figura 3.26.

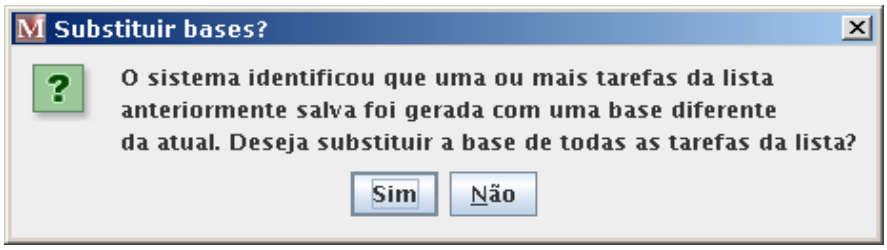
Substituir bases?
O sistema identificou que uma ou mais tarefas da lista anteriormente salva foi gerada com uma base diferente da atual. Deseja substituir a base de todas as tarefas da lista?
Sim
Não

Figura 3.26: Abrindo configuração - sistema questiona se é necessário substituir as bases de dados das tarefas

Caso o usuário decida substituir as bases, a lista será apresentada com todas as tarefas configuradas para nova base. Caso ele não queria substituir a base, a lista será apresentada conforme foi criada, porém as tarefas no qual o arquivo da base não mais existir em disco serão eliminadas.

3.2.6 Técnicas de classificação e estimativa de erro disponíveis e seus parâmetros configuráveis

Nesta seção serão apresentadas as técnicas de classificação e de estimativa de erro disponíveis no sistema. Para cada técnica será apresentada uma breve explicação do significado de cada parâmetro disponível.

Técnica de classificação *KNN* ou K vizinhos mais próximos

O algoritmo KNN, basicamente, filtra os k-vizinhos mais próximos de um dado exemplo. Para cada exemplo da base de consulta ou de validação, esse algoritmo varre a base de treinamento e procura os K exemplos que possuem a menor distância (os que estão mais próximos) do caso desconhecido. A distância, neste caso, é a distância Euclidiana entre os exemplos, onde os valores dos atributos da base são tidos como componentes de um vetor multidimensional.

Ao final da varredura, a técnica terá um vetor de tamanho K com os K vizinhos mais próximos. O critério de decisão é: dos K vizinhos mais próximos quantas vezes aparecem cada classe. A classe mais freqüente é a classe escolhida.

Na figura 3.27 são exibidos os parâmetros da técnica.

A seguir, cada parâmetro é brevemente descrito.

- Número de vizinhos - número dos K vizinhos a serem pesquisados. Assume valores inteiros positivos.

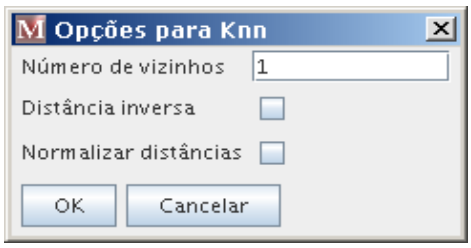

Figura 3.27: Parâmetros da técnica *KNN*

- Distância inversa - se marcada, a classe final do exemplo não será dada apenas pela classe mais freqüente entre os K vizinhos. Cada vizinho terá um peso atribuído (dado pelo inverso da sua distância) e a classe que obtiver a maior soma dos pesos será atribuída ao exemplo. Assume os valores Sim ou Não.

- Normalizar distância - normaliza as distâncias para o cálculo do peso de cada vizinho.

Técnica de classificação *KStar*

K* é um classificador baseado em instância. Seu funcionamento é muito semelhante ao do classificador Knn, ou seja, para cada instância da base de consulta ou de validação, esse classificador varre a base de treinamento e procura as K instâncias que mais se assemelham à instância desconhecida. Ao final da varredura, a classe apontada é aquela que apareceu com mais freqüência entre as instâncias semelhantes.

Na figura 3.28 são exibidos os parâmetros da técnica.

Figura 3.28: Parâmetros da técnica *KStar*

A seguir, cada parâmetro é brevemente descrito.

- Fator de mistura global - informa como a função de distância utilizada para calcular as k instâncias mais próximas irá agir. Quanto mais próximos de 0 forem os valores, menos exemplos serão considerados. Quanto mais próximos de 100 forem os valores, mais exemplos semelhantes serão utilizados para votar na classe mais freqüente. O aumento do fator global de mistura pode amenizar o efeito de instâncias ruidosas na classificação. Valores menores que zero, maiores que 100 ou não inteiros são proibidos.

- Mistura baseada em entropia - se marcado, a escolha das instâncias semelhantes será automática e guiada por critérios baseados em Entropia. Se não, seguirá o fator de mistura global desejado.

- Tratamento de nulos - informa qual o comportamento do classificador perante atributos com valores desconhecidos. Pode-se simplesmente ignorar as instâncias que os possuem, ou

então pode-se substituir o valor desconhecido pelo valor que maximiza a diferença entre as instâncias. Pode-se substituir, ainda, pelo valor médio do atributo, ou então considerar o valor desconhecido como um valor normal (um valor que pertence ao conjunto de valores possíveis para o atributo). Por exemplo, se os valores admitidos são ”Quente”, ”Frio”e ”Morno”, o novo conjunto seria ”Quente”, ”Frio”, ”Morno”e ”Normal”.

Técnica de classificação *Naive-Bayes*

Utiliza extensivamente a idéia de que os atributos são condicionalmente independentes dentro da classe. Embora esta suposição seja obviamente problemática, *Naive-Bayes* pode surpreender apresentando-se melhor do que muitos sofisticados classificadores em conjunto de dados nos quais as características não são fortemente combinadas.

O Naive Bayes não possui parâmetros.

Técnica de classificação *BayesNetB*

Classificador baseado em um algoritmo de busca por melhoria (*Hill Climbing Algorithm*) sem a restrição de ordem de variáveis.

Na figura 3.29 são exibidos os parâmetros da técnica.

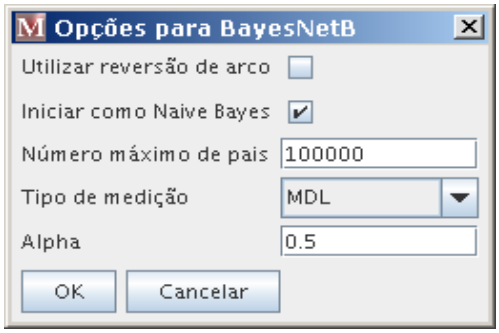

Figura 3.29: Parâmetros da técnica *BayesNetB*

A seguir, cada parâmetro é brevemente descrito.

- Utilizar reversão de arco - Se marcado, arcos reversos são considerados no próximo passo a ser feito.
- Iniciar como *Naive Bayes* - quando marcada, a rede inicial utilizada para a estrutura de aprendizagem é uma rede *Naive Bayes*. Se a opção não estiver marcada, uma rede vazia é usada como estrutura inicial.
- Número máximo de pais - limita o número máximo de pais que um nó de uma rede bayesiana pode ter.
- Tipo de medição - define o tipo de métrica que irá avaliar a qualidade da estrutura da rede (MDL/AIC).
- Alpha - parâmetro usado para a estimativa das tabelas de probabilidade do algoritmo.

Técnica de classificação *BayesNetk2*

Classificador baseado no *K2* para sua estrutura de aprendizagem. *K2* é um algoritmo de busca por melhoria (*Hill Climbing Algorithm*) criado por Greg Cooper e Ed Herskovitz. Esse algoritmo restringe a ordem das variáveis.

Na figura 3.30 são exibidos os parâmetros da técnica.

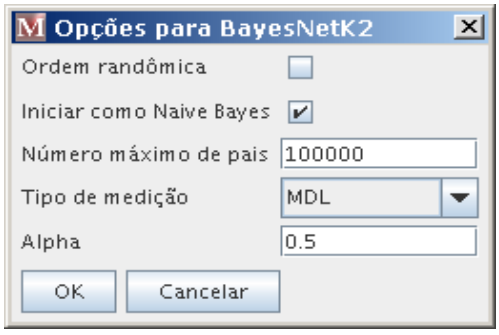

Figura 3.30: Parâmetros da técnica *BayesNetk2*

A seguir, cada parâmetro é brevemente descrito.

- Ordem randômica - quando marcada, a ordem dos nós na rede é randômica. Se não estiver marcada, a ordem dos nós no conjunto de dados é utilizada. Assume os valores Sim ou Não.
- Os demais parâmetros são os mesmo da técnica *BayesNetB*.

Técnica de classificação *J48*

É uma versão modificada do algoritmo C4.5 (Quinlan, 93) que descobre conhecimento na forma de árvore de decisão. Uma árvore de decisão é um diagrama de fluxo em estrutura de árvore, em que cada nó interno indica um teste em um atributo, cada ramificação representa um resultado de um teste e nós folha representam classes ou distribuições de classes.

Na figura 3.31 são exibidos os parâmetros da técnica.

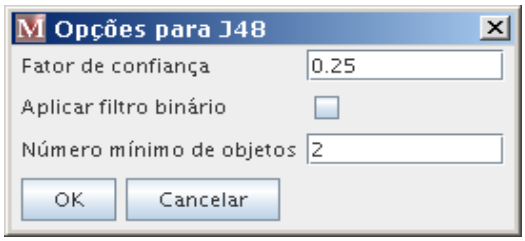

Figura 3.31: Parâmetros da técnica *J48*

A seguir, cada parâmetros é brevemente descrito.

- Fator de confiança - determina o valor de confiança a ser usado quando a poda da árvore é feita (valores menores acarretam maior poda). Assume valores entre 0 e 1.
- Número mínimo de objetos - número mínimo de instâncias que devem estar presentes no conjunto de treinamento para que uma nova folha seja criada em uma árvore de decisão. Assume valores inteiros positivos.

- Aplicar filtro binário - se marcado, cada valor possível de um atributo nominal gera uma novo atributo binário. Assume os valores Sim ou Não.

Técnica de classificação *NNge*

Algoritmo do tipo *vizinhos mais próximo*, utilizando exemplares generalizados não aninhados (hiperretângulos que podem ser vistos como regras do tipo *se...então*).

O NNge não possui parâmetros configuráveis.

Técnica de classificação *Perceptron Multi-Camada*

Baseado em células independentes de processamento (neurônios), o classificador *Perceptron Multi-Camada* consiste de múltiplas camadas de unidades computacionais, usualmente interconectadas de maneira que cada neurônio de uma camada tem conexões diretas com os neurônios da camada seguinte.

Na figura 3.32 são exibidos os parâmetros da técnica.

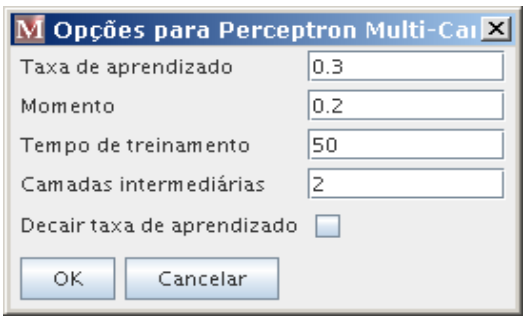

Figura 3.32: Parâmetros da técnica *Perceptron Multi-Camada*

A seguir, cada parâmetro é brevemente descrito.

- Taxa de aprendizado - a taxa com que uma nova instância influenciará na atualização dos pesos da rede. Assume valores entre 0 e 1.

- Momento - grau de inércia da rede, ou seja, a taxa com que o valor atual de cada peso influencia na sua atualização. Assume valores entre 0 e 1.

- Tempo de treinamento - número máximo de vezes em que todos os exemplos são aplicados durante o treinamento da rede. Assume valores inteiros positivos.

- Camadas intermediárias - define como será construída a estrutura intermediária da rede, que deve ser representada por uma lista de valores positivos separados por vírgula, um para cada camada. Cada valor representa o número de neurônios que aquela camada terá. Assume valores inteiros positivos.

- Decair taxa de aprendizado - se marcado, diminui gradativamente a taxa de aprendizado pela razão entre o valor inicial da taxa de aprendizado e o tempo de treinamento. Assume os valores Sim ou Não.

Técnica de classificação *Comparador de Curvas*

Comparador de Curvas é um classificador baseado na distância entre a curva de consumo medido do cliente e a curva de consumo típico da classe na qual o cliente se encaixa.

A distância pode ser gerada por qualquer método de diferenciação: Distância Euclidiana, Correlação Linear, etc.

A construção do classificador registra as medidas estatísticas das distâncias para a classe de consumidores normais e consumidores fraudadores.

Com esses dados estatísticos em mãos, é gerada uma distância de separação: os clientes que a ultrapassarem serão classificados como fraudadores; os que tiverem distâncias menores serão indicados como não-fraudadores.

Na figura 3.33 é exibido o parâmetro da técnica.

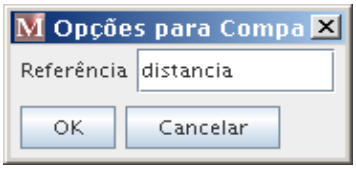

Figura 3.33: Parâmetro da técnica *Comparador de Curvas*

A seguir, o parâmetro é descrito.

- Referência - nome do atributo das instâncias (coluna da base) que guarda o valor da distância entre o consumo medido e o consumo esperado para o cliente.

Técnica de estimativa de erro *Validação Cruzada*

Divide o conjunto de dados em N sub-conjuntos e utiliza N-1 destes conjuntos para treinamento e o conjunto restante para teste de validação. Esta técnica repete este procedimento N vezes, cada vez utilizando um conjunto diferente para teste.

Na figura 3.34 é exibido o parâmetro da técnica.

Figura 3.34: Parâmetro da técnica *Validação Cruzada*

A seguir, o parâmetro é descrito.

- Número de conjunto - total de conjuntos criados. Deve, necessariamente, ser menor que o total de exemplos da base, caso contrário não será possível criar os conjuntos com elementos distintos.

Técnica de estimativa de erro *Divisão Percentual*

Divide o conjunto de dados em dois sub-conjuntos: um para treinamento e um para teste.

Na figura 3.35 é exibidos o parâmetro da técnica.

A seguir, o parâmetro é descrito.

Figura 3.35: Parâmetro da técnica *Divisão Percentual*

- Percentual de teste - define qual percentual da base original será reservada para teste. A divisão da base é feita aleatoriamente.

Técnica de estimativa de erro *Usar Conjunto de Treinamento*

Utiliza o mesmo conjunto de dados de treinamento para testar o desempenho do classificador. Não possui parâmetros de configuração. Esta técnica de validação deve ser usada com cuidado. A utilização da mesma base de treinamento como base de validação pode resultar em estatísticas de desempenho muito otimistas, mas que não serão obtidos novamente no Módulo de Consultas.

Técnica de estimativa de erro *Usar arquivo de teste*

Permite que um novo arquivo de dados seja selecionado para validação do classificador. O arquivo selecionado necessariamente deve estar em formato CSV idêntico ao arquivo de treinamento. O arquivo de teste será pré-processado da mesma forma que o arquivo de treinamento.

Na figura 3.36 é exibido o parâmetro da técnica.

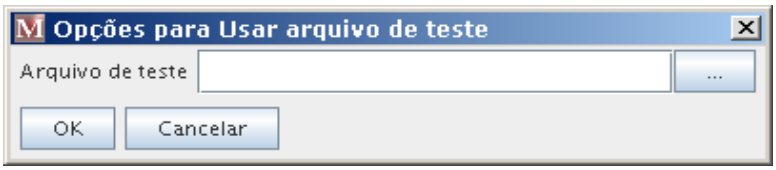

Figura 3.36: Parâmetro da técnica *Usar arquivo de teste*

A seguir, o parâmetro é descrito.

- Arquivo de teste - nome e caminho do arquivo CSV a ser utilizado na validação. Este parâmetro é obrigatório e o sistema não permitirá que a tarefa seja criada enquanto ele não estiver preenchido.

3.2.7 Busca por parâmetros de uma técnica de classificação

A busca por parâmetros utiliza um algoritmo genético que realiza os seguintes passos:

1. Inicialmente, cria-se uma população do tamanho escolhido. Uma população é um conjunto de cromossomos, o número de cromossomos da população é o tamanho da população. Um cromossomo representa uma configuração para a técnica escolhida. Na população inicial, as configurações são geradas aleatoriamente escolhendo-se uma dentre as possíveis, dados os valores que cada parâmetro pode assumir.

2. Essa população inicial passa por vários processos “evolutivos” para gerar uma nova população (nova geração). Esses processos são:

Mutação: um número de cromossomos dado por uma taxa multiplicada pelo tamanho da população é escolhido para sofrer mutação. Os cromossomos que obtiveram melhor desempenho têm maior probabilidade de serem escolhidos. A mutação consiste em escolher um parâmetro dentre os que compõem a configuração para ter seu valor modificado. O parâmetro que sofrerá a modificação e o novo valor são escolhidos aleatoriamente. Os cromossomos que sofrerem mutação farão parte da nova população.

Crossover: um número de cromossomos dado por uma taxa multiplicada pelo tamanho da população é escolhido para sofrer crossover. Os cromossomos que obtiveram melhor desempenho têm maior probabilidade de serem escolhidos. O crossover é realizado em pares de cromossomos. Ele consiste em escolher um parâmetro e a partir dele todos os parâmetros terão seus valores trocados: os valores do cromossomo1 passarão para o cromossomo2 e os valores do cromossomo2 passarão para o cromossomo1. Esse parâmetro é escolhido aleatoriamente. Os cromossomos que sofrerem crossover farão parte da nova população.

Elitismo: os cromossomos que obtiveram o melhor desempenho são selecionados para compor a nova população. O número de cromossomos selecionados no elitismo será o número de cromossomos que, somado ao número de cromossomos selecionados para mutação e ao número de cromossomos selecionados para crossover, completará o tamanho da população.

3. Este processo de geração de uma nova população se repetirá por um número de vezes pré-determinado, ou até que todos os cromossomos da população gerada tenham o mesmo desempenho. Após a última geração ter sido criada, o cromossomo de melhor desempenho nessa geração representará a configuração escolhida pelo método de busca.

Na figura 3.37 são exibidos os parâmetros configuráveis, do método de busca.

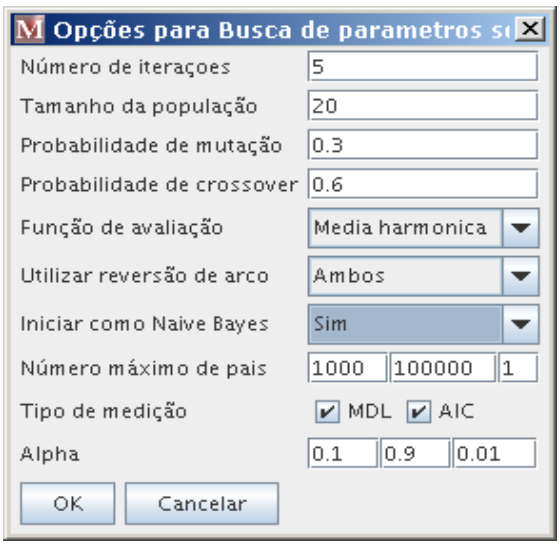

Figura 3.37: Opções da busca com a técnica *BayesNetB*

A seguir, os parâmetros do método de busca são descritos.

- Número de iterações - é o número de vezes que uma nova população será criada. Deve ser um número inteiro positivo.
- Tamanho da população - é o número de cromossomos que a população terá, em todas as gerações. Deve ser um número inteiro positivo.

- Probabilidade de mutação - é o percentual de cromossomos da população que sofrerá mutação para compor a nova população. Deve ser um valor real entre 0 e 1.

- Probabilidade de crossover - é o percentual de cromossomos da população que sofrerá crossover para compor a nova população. Deve ser um valor real entre 0 e 1. A soma das probabilidades de mutação e crossover deve ser um valor real entre zero e um. A diferença entre 1 e esse valor será o percentual de cromossomos escolhido por elitismo para formar o restante dos cromossomos da nova população.

- Função de avaliação - é a função utilizada para medir o desempenho do classificador gerado por uma configuração representada por meio de um cromossomo. Pode ser uma dentre as seguintes: média harmônica (entre especificidade e confiabilidade negativa), média geométrica (entre especificidade e confiabilidade negativa), acerto (taxa de acerto do classificador), amplitude (especificidade) ou precisão (confiabilidade negativa).

Além desses parâmetros, para cada técnica devem ser escolhidos os valores que cada parâmetro seu poderá assumir para formar o conjunto de configurações possíveis para essa técnica.

Na figura 3.37 temos o exemplo de uma tela para configuração das opções da busca para a técnica BayesNetB. As cinco primeiras opções são iguais para qualquer técnica (são os parâmetros para a configuração de qualquer busca). As outras opções permitem configurar os valores que cada parâmetro da técnica podem assumir. No exemplo, os parâmetros e os valores que podem assumir são: utilizar reversão de arco: ambos (assume os valores sim e não); iniciar como *Naive Bayes*: só assume o valor sim; número máximo de pais: assume valores entre 1000 e 100000 com variação de 1; tipo de medição: assume os valores MDL e AIC; alpha: assume valores entre 0.1 e 0.9 com variação de 0.01.

Capítulo 4

Módulo de Consultas

O Módulo de Consultas é a parte do sistema que permite ao usuário realizar a classificação de um conjunto de consumidores utilizando os classificadores criados anteriormente através do Módulo de Aprendizagem. Esse é o módulo que gera o produto final do software: exemplos selecionados para inspeção.

4.1 Conhecendo as interfaces do Módulo de Consultas

Nesta seção serão apresentadas as interfaces do Módulo de Consultas e será descrito como cada uma delas é utilizada.

4.1.1 Janela inicial

A janela principal, apresentada na figura 4.1, permite o acesso às diversas funcionalidades do sistema, descritas posteriormente.

4.1.2 Menu Arquivo

O Menu “Arquivo”, exibido na figura 4.2, possui as seguintes opções:

- “Nova configuração” - permite criar uma nova configuração.
- “Abrir configuração” - permite que uma configuração existente seja aberta.
- “Salvar configuração” - permite salvar a configuração atual.
- “Salvar como” - permite salvar a configuração atual com outro nome desejado.
- “Sair” - permite ao usuário sair do Módulo de Consultas.

Uma configuração no Módulo de Consultas corresponde a uma lista de classificadores ao modo de avaliação selecionado para consulta.

As mesmas funções disponíveis no menu arquivo também podem ser acionadas através da barra de botões da tela principal do sistema. Esta barra de botões é ilustrada na figura 4.3.

4.1.3 Seleção da base de dados

Assim como no Módulo de Aprendizagem, na parte superior da tela principal do Módulo de Consultas existe um painel para seleção e manipulação da base de exemplos para consulta. Neste

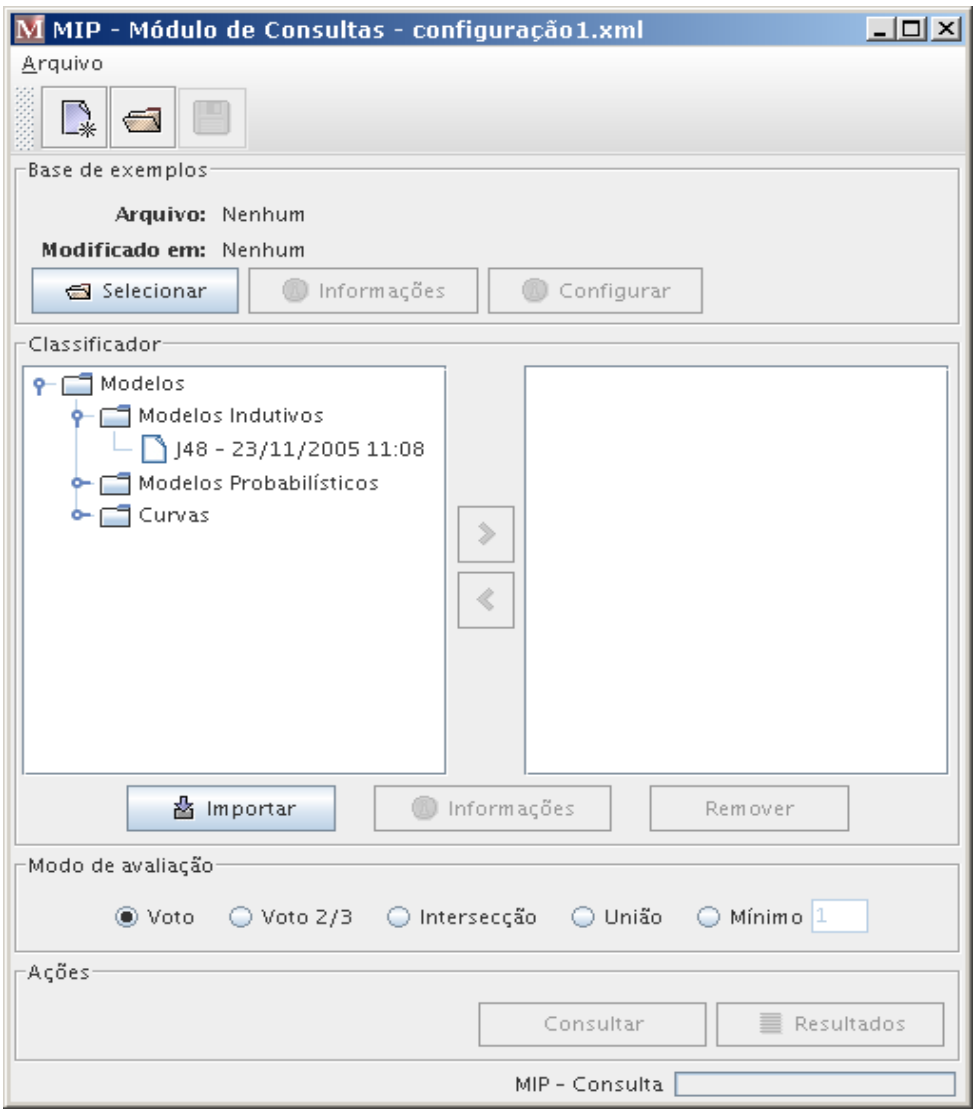

Figura 4.1: Módulo de consultas - tela principal

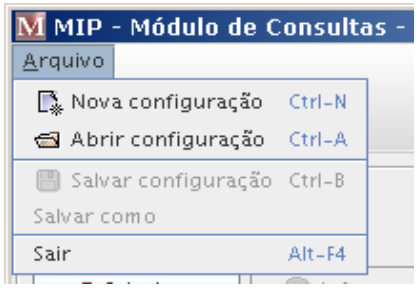

Figura 4.2: Módulo de consultas - menu arquivo

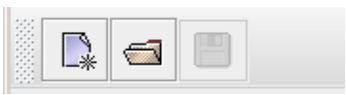

Figura 4.3: Módulo de consultas - botões de configuração

painel é possível selecionar, configurar e obter informações a respeito da base de dados. Na figura 4.4 este painel é ilustrado.

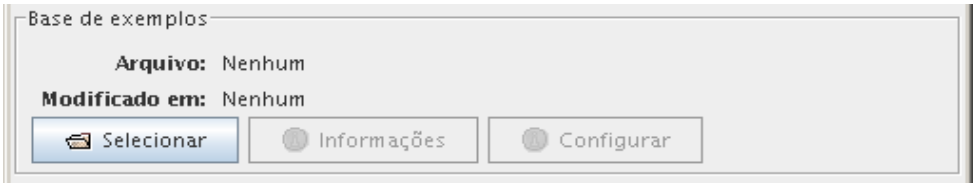

Figura 4.4: Painel de seleção de base de consulta

O funcionamento desta parte do sistema é descrito em detalhes no capítulo 5.

4.1.4 Manipulação e seleção de técnicas

No painel de seleção de classificadores estão localizados dois botões representados por "setas" que permitem a construção de uma lista de classificadores que serão aplicados à base escolhida. Para adicionar um classificador à lista, basta selecioná-lo na árvore à esquerda e então clicar no botão "seta à direita". O classificador selecionado será, então, inserido na lista de classificadores à direita. Para remover um classificador da lista de selecionados, basta clicar no botão "seta à esquerda".

Neste mesmo painel encontram-se os botões "Importar", "Remover" e "Informações". O botão "Importar" permite realizar a importação de um classificador exportado no Módulo de Aprendizagem. O botão "Remover" remove da árvore de classificadores o classificador que estiver selecionado. O botão "Informações" exibe informações a respeito de um classificador selecionado na árvore à esquerda do painel.

O painel de manipulação de técnicas é exibido na figura 4.5.

Figura 4.5: Painel de seleção de classificadores

4.1.5 Modos de avaliação

O painel de configuração do modo de avaliação permite a seleção do modo de decisão que será considerado para classificação de um caso desconhecido entre normal e fraudador, quando mais de uma técnica estiver sendo usada na avaliação.

Atualmente estão disponíveis cinco formas de avaliação, que serão descritas em detalhes na seção 4.2.2. Na figura 4.6 o painel de seleção do modo de avaliação é exibido.

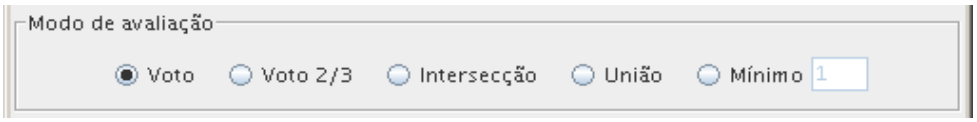

Figura 4.6: Painel de seleção de modo de avaliação

4.1.6 Botões de controle e barra de estado

Abaixo do painel de configuração do modo de avaliação encontram-se os botões “Consultar” e “Resultados”. O botão “Consultar” inicia o processo de consulta, aplicando os classificadores selecionados à base escolhida de acordo com o modo de avaliação corrente. Durante a execução de uma consulta o botão “Consultar” transforma-se em “Cancelar”. O acionamento do botão “Cancelar” pára o processamento da consulta e nenhum resultado é gerado.

O botão “Resultado” permite o acesso à tela de exibição de resultados da consulta. Somente o último resultado é mostrado. Esta parte da tela principal é ilustrada na figura 4.7.

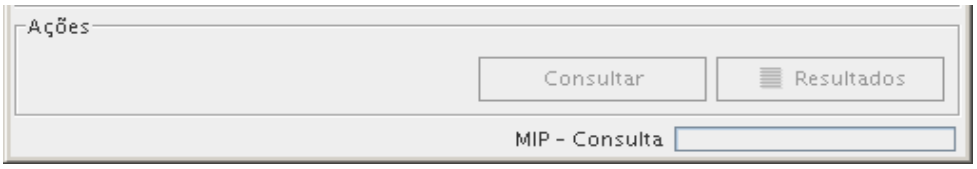

Figura 4.7: Botões “Consultar” e Resultado”

Abaixo dos botões de controle está a barra de estado que exibe mensagens informativas durante o processo de consulta e de seleção de uma base de exemplos. Durante a execução de uma consulta, a barra de estado entra em modo de execução para indicar que o processo está em andamento.

4.1.7 Janela de resultados

Após realizar uma consulta, estará disponível para o usuário a lista de todos os consumidores avaliados e a classificação obtida. Na janela de resultados é possível verificar o código do usuário e a classificação obtida. Na mesma tela, ilustrada na figura 4.8, também é possível salvar o resultado em formato CSV.

As funcionalidades desta tela são descritas em detalhes na seção 4.2.3.

4.2 Funcionalidades do Módulo de Consultas

Nesta seção são descritas em detalhes as funcionalidades do Módulo de Consultas.

4.2.1 Importando e removendo um classificador

Antes que um classificador treinado esteja disponível para ser utilizado em operações de consulta, é necessário que ele seja importado. Essa opção irá incluir o classificador selecionado na árvore de classificadores disponíveis.

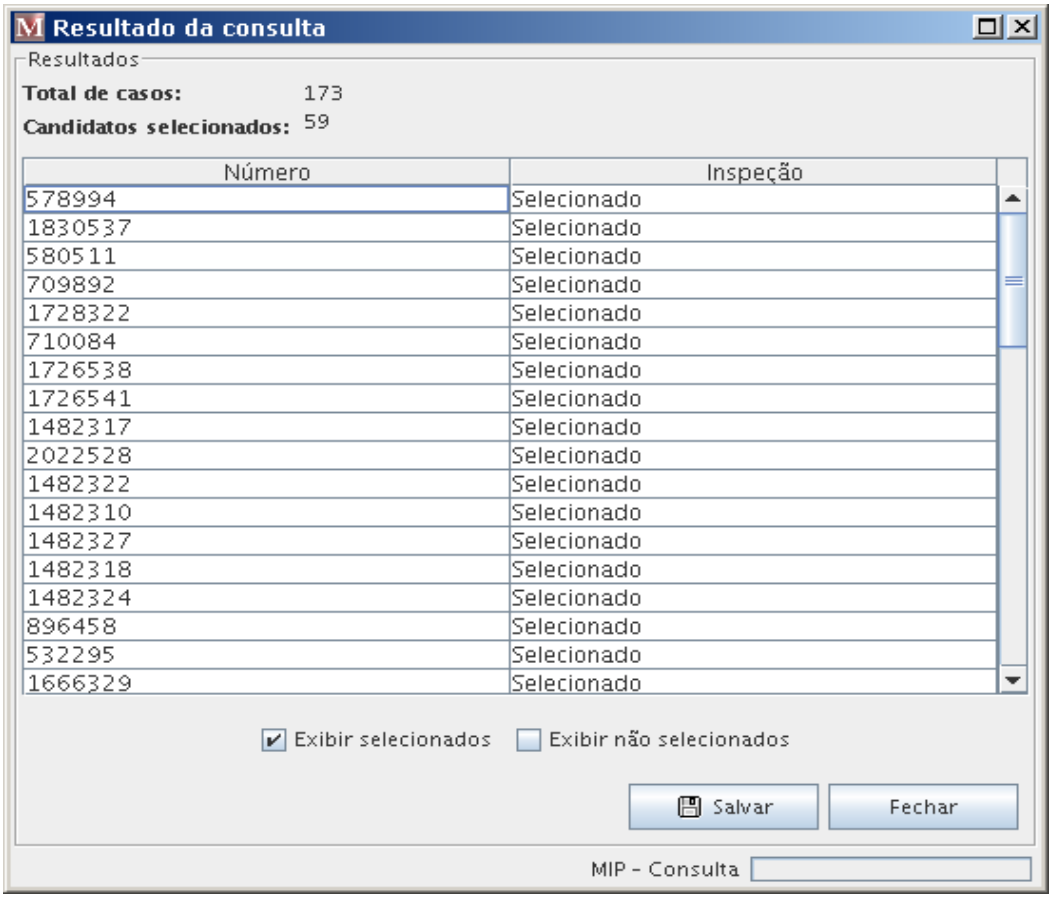

Figura 4.8: Tela de exibição de resultados

Para realizar a importação, pode-se clicar com o botão direito na árvore de classificadores ou no botão “Importar” localizado abaixo da árvore de classificadores, como mostrado na figura 4.9.

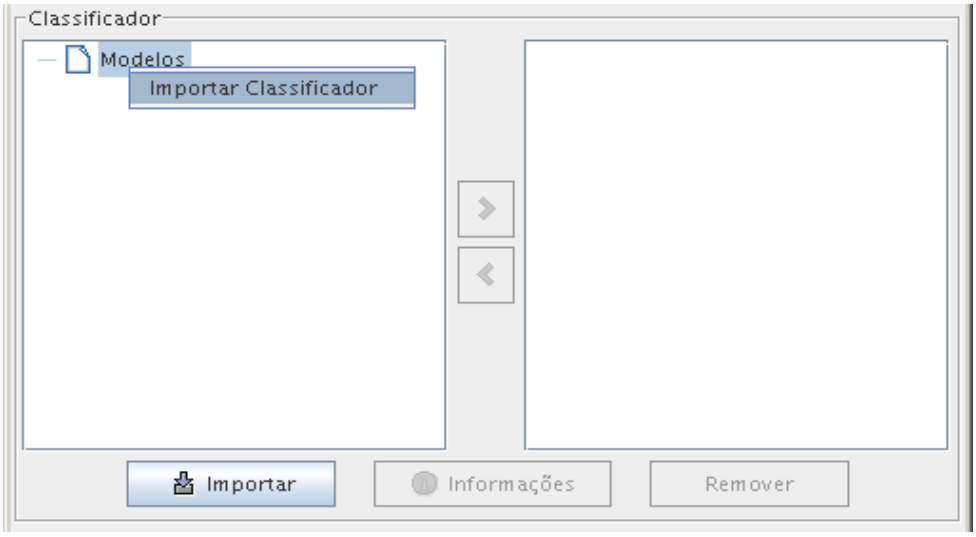

Figura 4.9: Importando classificadores

Ao clicar-se em “Importar Classificador”, aparecerá uma caixa de diálogo para que seja selecionado o classificador desejado.

Após terminar a importação, o classificador selecionado será incluído na árvore de classificadores, como mostrado na figura 4.10.

Figura 4.10: Classificadores importados

Ao selecionar um classificador importado, são habilitados os botões “Remover” e “Informações”. Clicando-se no botão “Remover” o classificador selecionado é removido da árvore de classificadores - o que também pode ser feito clicando-se com o botão direito sobre o classificador, como mostrado na figura 4.11.

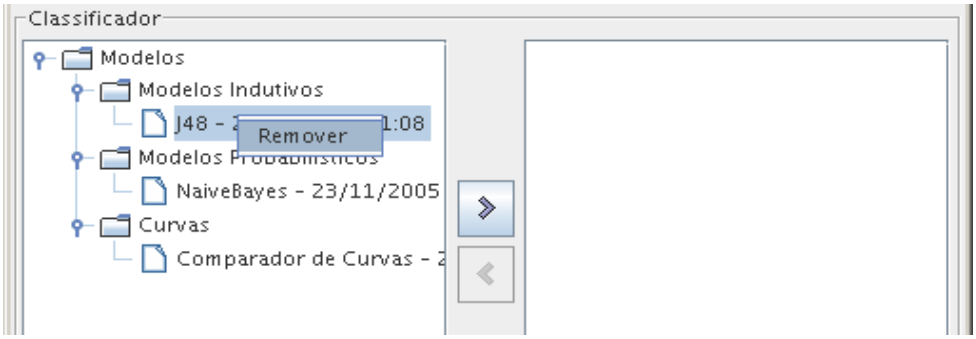

Figura 4.11: Removendo um classificador

Ao selecionar um classificador da árvore de classificadores e, em seguida, acionar o botão “Informações”, serão exibidas informações a respeito do treinamento do classificador selecionado: parâmetros, estimador de erro e número de instâncias utilizadas no treino, estimativas de erro etc. Um exemplo disto é mostrado na figura 4.12.

4.2.2 Realizando uma consulta

Os seguintes passos são necessários à realização de uma consulta:

1. Escolher base de casos: inicialmente é necessário selecionar um arquivo que conterá o conjunto de consumidores que se deseja classificar. A seleção de uma base de caso deve ser feita de acordo com as instruções do capítulo 5.

2. Selecionar classificadores: após selecionar a base de casos, é necessário selecionar as técnicas para formação da lista de classificadores (um ou mais) escolhidos.

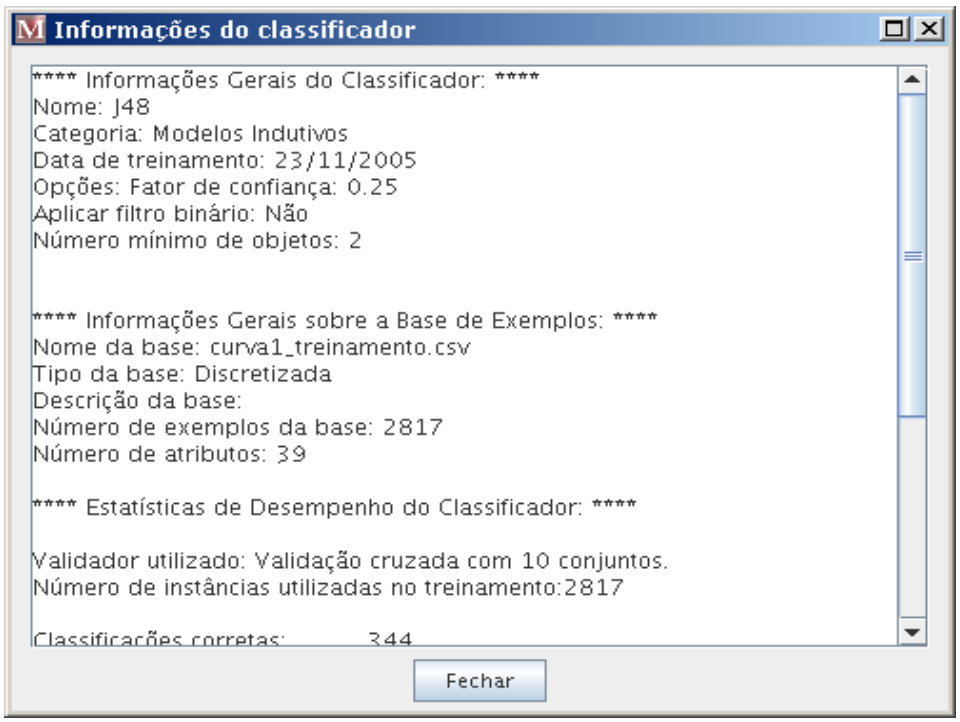
Informações do classificador
**** Informações Gerais do Classificador: ****
Nome: J48
Categoria: Modelos Indutivos
Data de treinamento: 23/11/2005
Opções: Fator de confiança: 0.25
Aplicar filtro binário: Não
Número mínimo de objetos: 2
**** Informações Gerais sobre a Base de Exemplos: ****
Nome da base: curva1_treinamento.csv
Tipo da base: Discretizada
Descrição da base:
Número de exemplos da base: 2817
Número de atributos: 39
**** Estatísticas de Desempenho do Classificador: ****
Validador utilizado: Validação cruzada com 10 conjuntos.
Número de instâncias utilizadas no treinamento:2817
Classificações corretas: 344
Fechar

Figura 4.12: Visualizando informações de um classificador

Para incluir um classificador na lista, deve-se clicar no classificador e em seguida no botão "seta à direita". Feito isso, o classificador será adicionado na lista da direita, como mostrado na figura 4.13.

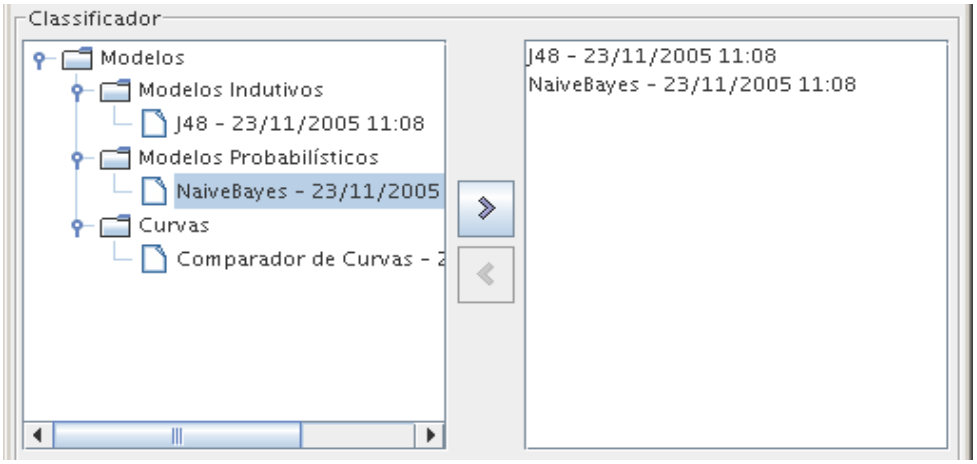

Figura 4.13: Selecionando classificadores

3. Escolher o modo de avaliação: existem cinco modos de avaliação disponíveis: voto, voto 2/3, intersecção, união e mínimo.

   A escolha de cada modo de avaliação fará com que o sistema mude a forma de decisão da classe dos exemplos desconhecidos no Módulo de Consultas.

   O modo de avaliação escolhido somente será considerado caso várias técnicas forem selecionadas para consulta simultânea. Caso apenas uma técnica tenha sido selecionada para consulta (no passo anterior), a classificação dos consumidores será exclusivamente decidida pela técnica.

Caso várias técnicas tenham sido selecionadas, o sistema fará a classificação dos exemplos usando cada técnica selecionada e decidirá a qual classe o exemplo pertence usando o modo de avaliação selecionado.

As regras de decisão implementadas por cada modo de avaliação podem ser assim descritas:

- Voto: serão classificados como selecionados para inspeção os consumidores que assim foram rotulados pela maioria das técnicas. Em caso de empate, o consumidor é rotulado como selecionado para inspeção.
- Voto 2/3: serão classificados como selecionados para inspeção os consumidores que assim foram rotulados por pelo menos dois terços das técnicas selecionadas.
- Intersecção: serão classificados como selecionados para inspeção os consumidores que assim foram rotulados por todas as técnicas selecionadas.
- União: serão classificados como selecionados para inspeção os consumidores que assim foram rotulados por pelo menos uma técnica.
- Mínimo: serão classificados como selecionados para inspeção os consumidores que assim foram rotulados por pelo menos um número de mínimo de técnicas. O mínimo necessário, neste caso, é passado como parâmetro na interface.

Na figura 4.14 tem-se uma demonstração da escolha do modo de avaliação “mínimo de 2 técnicas”.

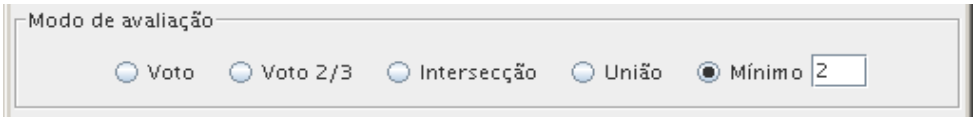

Figura 4.14: Escolhendo o modo de avaliação

4. Iniciar a classificação: após executar os três passos anteriores, o botão “Consultar” deve ser acionado para dar início à consulta. Durante o processo de consulta o botão “Consultar” é convertido em “Cancelar”. O acionamento do botão “Cancelar” pára a consulta imediatamente.

4.2.3 Verificando resultados

Após a realização bem-sucedida de uma consulta, é exibida a tela de visualização de resultados, como mostrado na figura 4.15. Se fechada, esta tela pode ser acessada novamente através do botão “Resultados” da tela principal.

O resultado com a classificação de todos os consumidores selecionados é exibido na tabela central da tela de resultados. Na parte superior da tela são disponibilizadas informações do número total de consumidores existentes na base de casos e do número de consumidores que foram classificados como selecionados.

Na parte inferior, é possível escolher entre a exibição ou não dos consumidores que foram classificados como selecionados, e o mesmo para os não selecionados.

Na figura 4.16 temos um exemplo das caixas de seleção, no qual foi escolhida a exibição apenas dos consumidores classificados como selecionados.

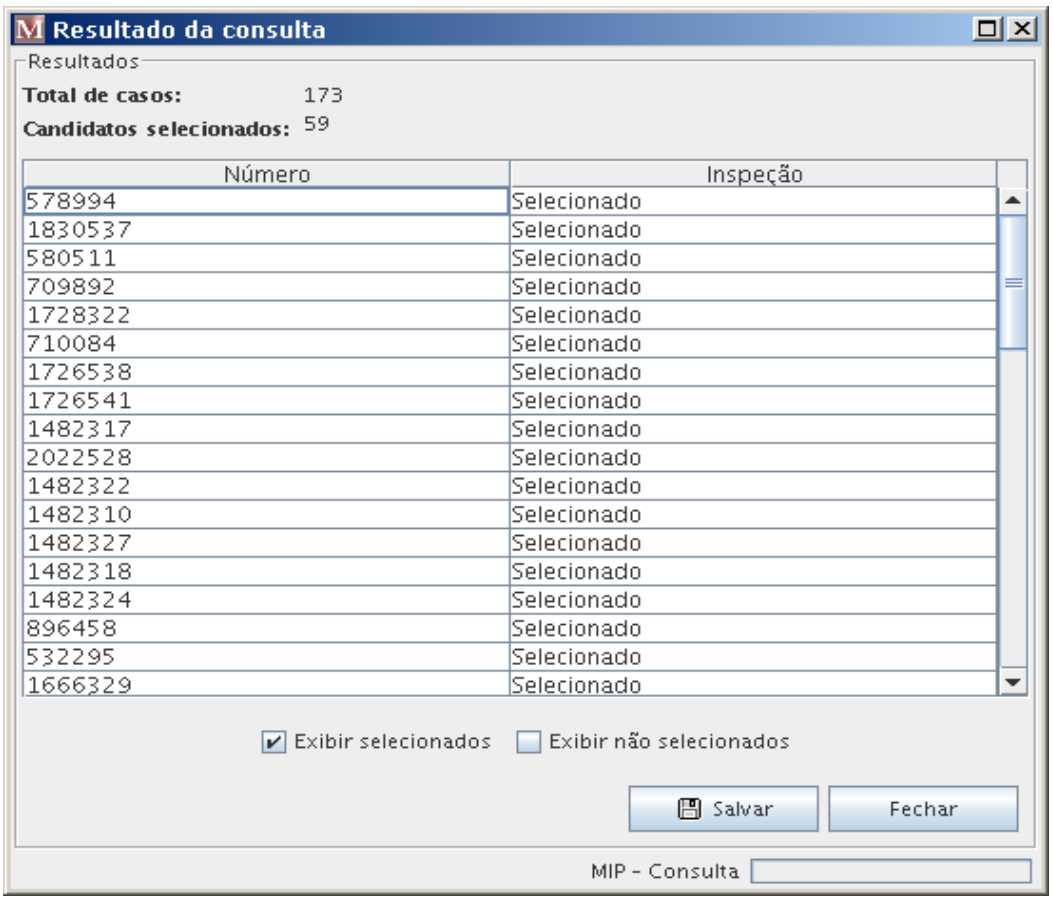

Figura 4.15: Tela de visualização de resultados

Figura 4.16: Exibindo apenas registros classificados como selecionados

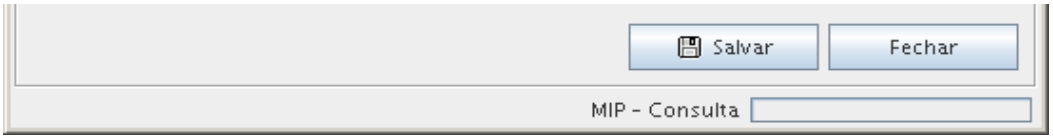

Figura 4.17: Botão para salvar o resultado em formato CSV

O botão “Salvar”, mostrado na figura 4.17, permite gravar em arquivo (no formato CSV) o conteúdo da tabela.

Para sair da tela de resultados deve-se clicar no botão “Fechar”.

4.2.4 Criando, recuperando e salvando configurações

Uma configuração de consultas consiste em um conjunto de várias técnicas selecionadas para consulta e do modo de avaliação corrente.

Como ilustração, suponha que tenha sido criada a configuração de consulta mostrada na figura 4.18.

Neste momento, o sistema mostrará que nenhuma configuração ainda foi salva ou aberta

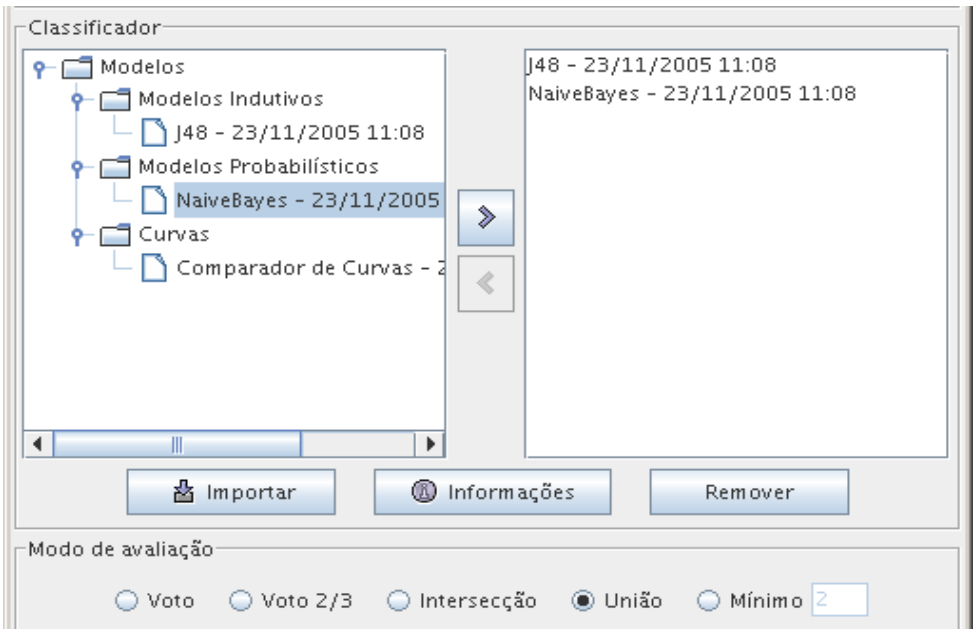

Figura 4.18: Tela com uma configuração para ser salva

através da mensagem na barra de título mostrada na figura 4.19.

Figura 4.19: Barra de título padrão - nada ainda foi salvo ou carregado

Usando os menus “Salvar configuração” ou “Salvar como” (ou ainda os botões da barra de tarefas mostrados na seção 4.1.2), é possível salvar a configuração atual. Por exemplo, acionando o menu “Salvar Configuração”, ilustrado na figura 4.20, será solicitado ao usuário o nome do arquivo de configuração a ser salvo.

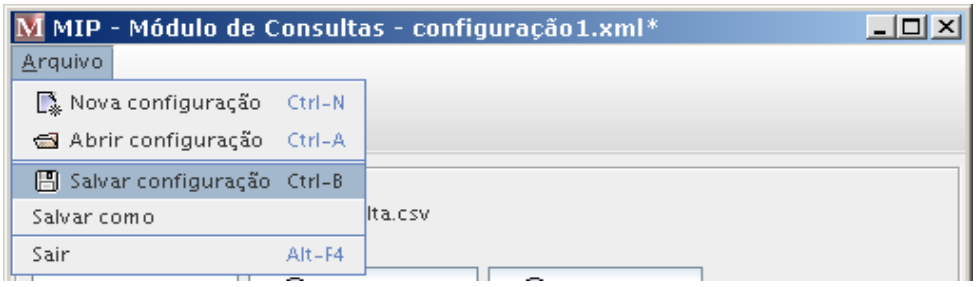

Figura 4.20: Menu “Salvar”

Após acionar o menu “Salvar configuração”, o sistema mostrará na barra de título o nome da configuração atualmente carregada, conforme ilustrado na figura 4.21. Notar que neste caso não existe o caractere * em frente ao nome da configuração. A ausência deste caractere indica que nenhuma alteração foi feita na configuração em relação ao que foi originalmente carregado do arquivo salvo em disco.

Caso o usuário decida iniciar uma nova configuração, usando a opção “Nova configuração”, a lista de técnicas selecionadas será esvaziada.

Ao abrir uma configuração salva, a lista de tarefas selecionadas será preenchida com todas as

Figura 4.21: Barra de título após salvar a configuração

técnicas anteriormente salvas que ainda permanecem na árvore de técnicas disponíveis e o modo de avaliação será marcado de acordo com o anteriomente salvo.

Capítulo 5

Preparando e selecionando dados

Os dois módulos do sistema, apresentados nos capítulos anteriores, possuem o mesmo painel para seleção da base de dados de exemplos. Neste capítulo são apresentadas as funcionalidades deste painel.

5.1 Selecionando bases de dados no sistema

A primeira funcionalidade do painel é a seleção de arquivos de dados. Na figura 5.1 é mostrado o painel de seleção da base de exemplos após o acionamento do botão “Selecionar”. A partir deste botão é possível selecionar dois tipos de arquivos: arquivos CSV (arquivos no formato de texto, com colunas separadas por vírgula) e arquivos no formato do Datawarehouse da ESCELSA. Nas próximas seções, as duas opções são descritas em detalhes.

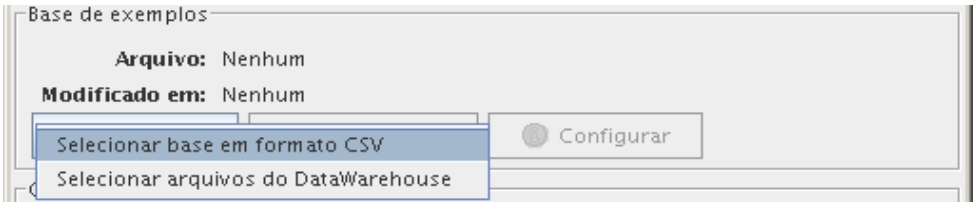

Figura 5.1: Selecionando uma base de dados no sistema

5.2 Selecionando uma base de dados em formato CSV

A primeira forma possível de carregar uma base de exemplos é selecionar uma base em formato CSV. Após clicar no botão “Selecionar” do painel de seleção da base de exemplos, será exibido o menu no qual é possível selecionar a base de exemplos em formato CSV (figura 5.1). Imediatamente após clicar no item de menu de seleção em formato CSV, será exibida a tela ilustrada na figura 5.2. Para abrir um novo arquivo basta selecioná-lo e clicar no botão “Abrir”. O nome do arquivo e a sua data de última modificação serão mostrados no painel de seleção de bases da tela principal do sistema.

Qualquer arquivo CSV válido pode ser usado no sistema MIP. Para que um arquivo texto seja um arquivo CSV válido, ele deve conter um cabeçalho com o nome das colunas do arquivo e uma ou mais linhas contendo os dados dos exemplos. Todas as linhas devem ter exatamente a mesma quantidade de colunas (ou seja, a mesma quantidade de separações por vírgula). Por exemplo,

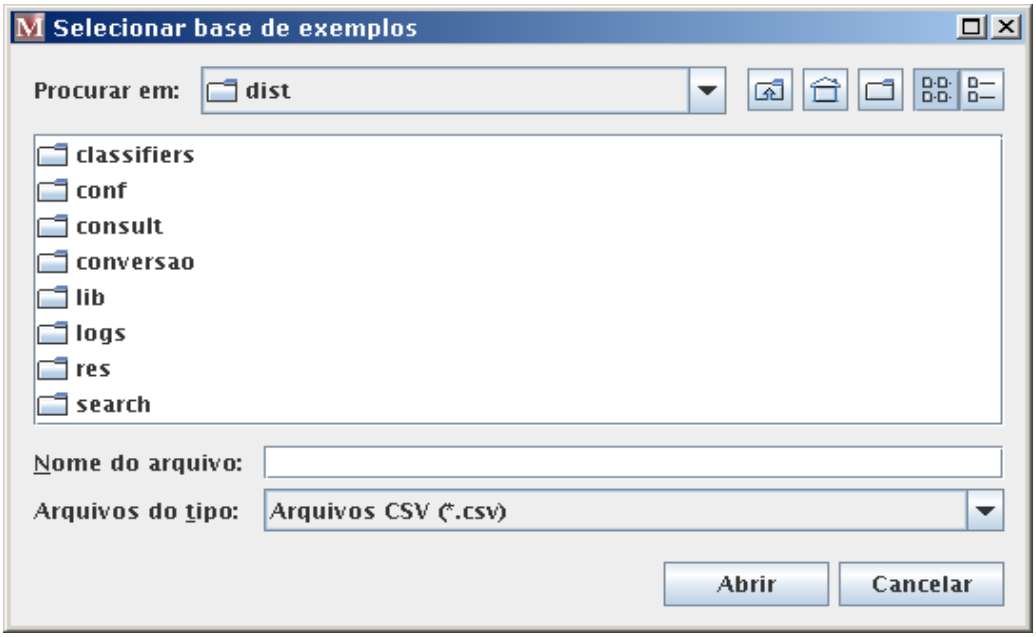

Figura 5.2: Selecionando uma base de dados no formato CSV

o texto abaixo é o conteúdo de um arquivo CSV válido. Neste exemplo, o arquivo contém seis colunas e dez exemplos.

```
dw_uc,dw_tarifa,media,desvio_padrao,mediatriz,resultado
1,2,100,120,120,1
3,2,120,180,130,1
4,2,130,120,140,0
5,2,140,150,150,0
6,2,100,150,160,0
7,6,160,110,180,0
8,5,140,110,130,1
32,4,120,110,140,1
34,null,140,110,150,0
46,2,160,140,120,1
```

Caso o arquivo deva conter números com separação de casas decimais, deve-se usar o ponto como separador (formato de separação americano).

O sistema MIP espera que o arquivo CSV atenda a quatro convenções:

1. A primeira coluna do arquivo será sempre um identificador único do exemplo em questão. No arquivo acima, a coluna dw_uc é um identificador único da base de dados.

2. A última coluna conterá a classe a qual o exemplo pertence. O sistema MIP também convenciona que esta coluna conterá apenas dois valores: 0 ou 1. Zero significa que o exemplo não é fraudador e um significa que o exemplo é fraudador.

3. O arquivo CSV conterá pelo menos uma coluna além das duas descritas nos itens anteriores.

4. Exemplos cujo valor de alguma coluna seja desconhecido, terá o valor preenchido com "null". No exemplo anterior, o valor da coluna "dw_tarifa" do penúltimo exemplo é desconhecido.

Os erros mais comuns na preparação da base de dados em formato CSV são listados a seguir. Espera-se que esta lista ajude na identificação da causa do problema quando o sistema exibir a mensagem de erro acusando a anomalia. Para cada problema, é exibida a mensagem de erro do sistema e descrito de quando esta será exibida.

1. O total de classes na base é diferente do esperado (2): neste caso, o sistema somente identificará que o total de classes está incorreto quando for realizar um novo treinamento (Módulo de Aprendizagem). No caso do Módulo de Consultas este erro nunca será exibido (veja seção 5.4.5). Após o acionamento do botão "Treinar" será exibida a mensagem da figura 5.3.

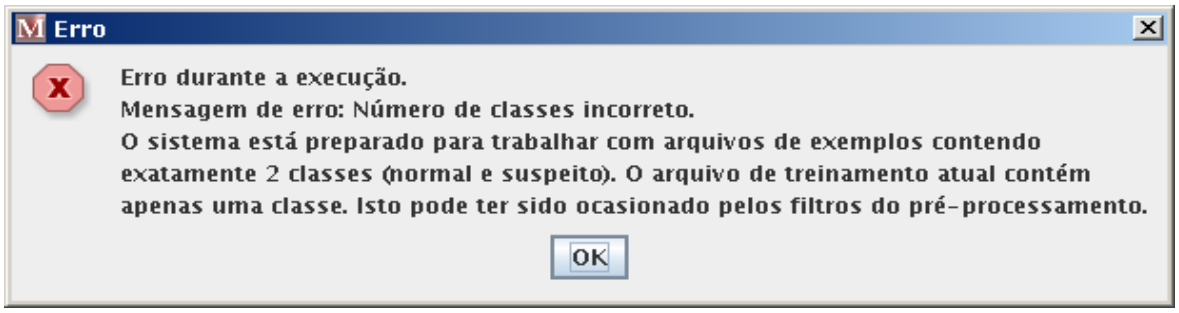
Erro
Erro durante a execução.
Mensagem de erro: Número de classes incorreto.
O sistema está preparado para trabalhar com arquivos de exemplos contendo exatamente 2 classes (normal e suspeito). O arquivo de treinamento atual contém apenas uma classe. Isto pode ter sido ocasionado pelos filtros do pré-processamento.
OK

Figura 5.3: Base selecionada contém número de classes incorreto

2. A base não contém cabeçalho de identificação das colunas: neste caso o sistema irá entender que a primeira linha contém o cabeçalho. Como resultado, as colunas da base terão como nome os dados da primeira linha do arquivo. O sistema exibirá mensagem de erro informando que a base está incompatível, conforme descrito na seção 5.5, e solicitará que a base seja reconfigurada. Neste caso recomenda-se que o usuário cancele a configuração automática do sistema e acerte a base antes de prosseguir (ver seção 5.5).

3. A base em formato CSV tem cabeçalho de identificação das colunas, mas o cabeçalho não lista todas as colunas do arquivo: neste caso o sistema identificará o problema logo na abertura do arquivo. Quando o cabeçalho da base de dados for diferente do cabeçalho da última base usada no MIP, o sistema reagirá conforme descrito na seção 5.5. Quando o cabeçalho for o mesmo da base usada na última seção do MIP (por exemplo, porque foi acrescentada uma coluna à base mas não foi corrigido o cabeçalho) será exibida a mensagem da figura 5.4.

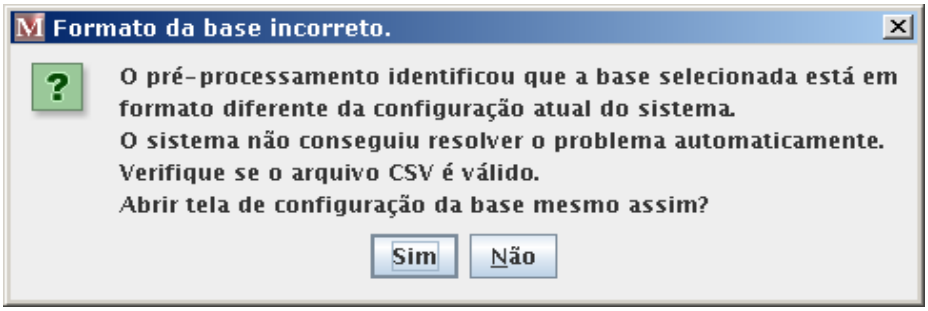
Formato da base incorreto.
O pré-processamento identificou que a base selecionada está em formato diferente da configuração atual do sistema.
O sistema não conseguiu resolver o problema automaticamente.
Verifique se o arquivo CSV é válido.
Abrir tela de configuração da base mesmo assim?
Sim
Não

Figura 5.4: O arquivo CSV é inválido - o sistema não sabe como resolver o problema

4. Alguma linha contém números com vírgula como separação de casas decimais: o sistema reagirá da mesma forma que no caso do item anterior.

5. As colunas da base CSV não estão separadas por vírgula: neste caso, o sistema entenderá que o arquivo contém apenas uma coluna.

5.3 Obtendo informações da base

Após selecionar uma base (tanto em formato CSV, conforme já descrito, quanto em formato do Datawarehouse, conforme será descrito na seção 5.6) o botão “Informações” do painel de seleção de bases é habilitado.

Este botão, quando acionado, exibe o total de exemplos da base e o total de atributos (colunas) que compõe a base, conforme ilustrado na figura 5.5.

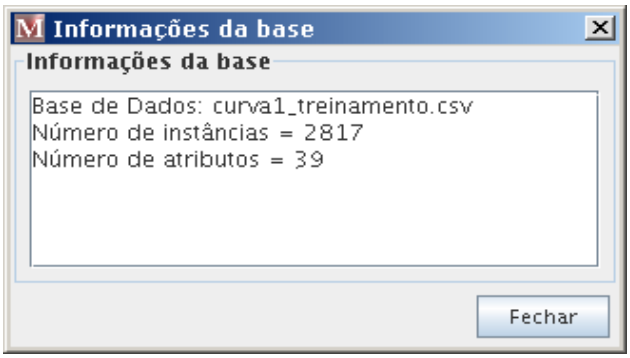

Figura 5.5: Obtendo informações da base selecionada

5.4 Configurando a base de dados

Conforme descrito, uma base de dados CSV é um conjunto de dados em formato texto que é apresentado ao sistema em um arquivo único, onde cada linha contém um exemplo e cada coluna representa um atributo. Conforme será mostrado na seção 5.6, a seleção de arquivos em formato do Datawarehouse também proporcionará ao MIP um arquivo neste formato - apenas ocorrerá um passo a mais de conversão de formato.

Do arquivo CSV, o sistema MIP só poderá obter uma única informação a respeito dos atributos (colunas) dos exemplos: os nomes dos atributos - obtidos a partir do nome das colunas do arquivo.

Como o sistema MIP necessita de mais informações a respeito das colunas dos arquivos para realização do treinamento e das consultas, existe uma interface de configuração da base de dados, que pode ser visualizada através do botão “Configurar” do painel de seleção de bases na tela principal dos módulos do sistema.

A tela de configuração da base de exemplos é ilustrada na figura 5.6. Nesta tela existem três grupos de configurações que podem ser alterados: descrição da base, gerenciador de colunas e gerenciador de filtros de descarte. Nas próximas seções, cada um desses grupos é descrito em detalhes.

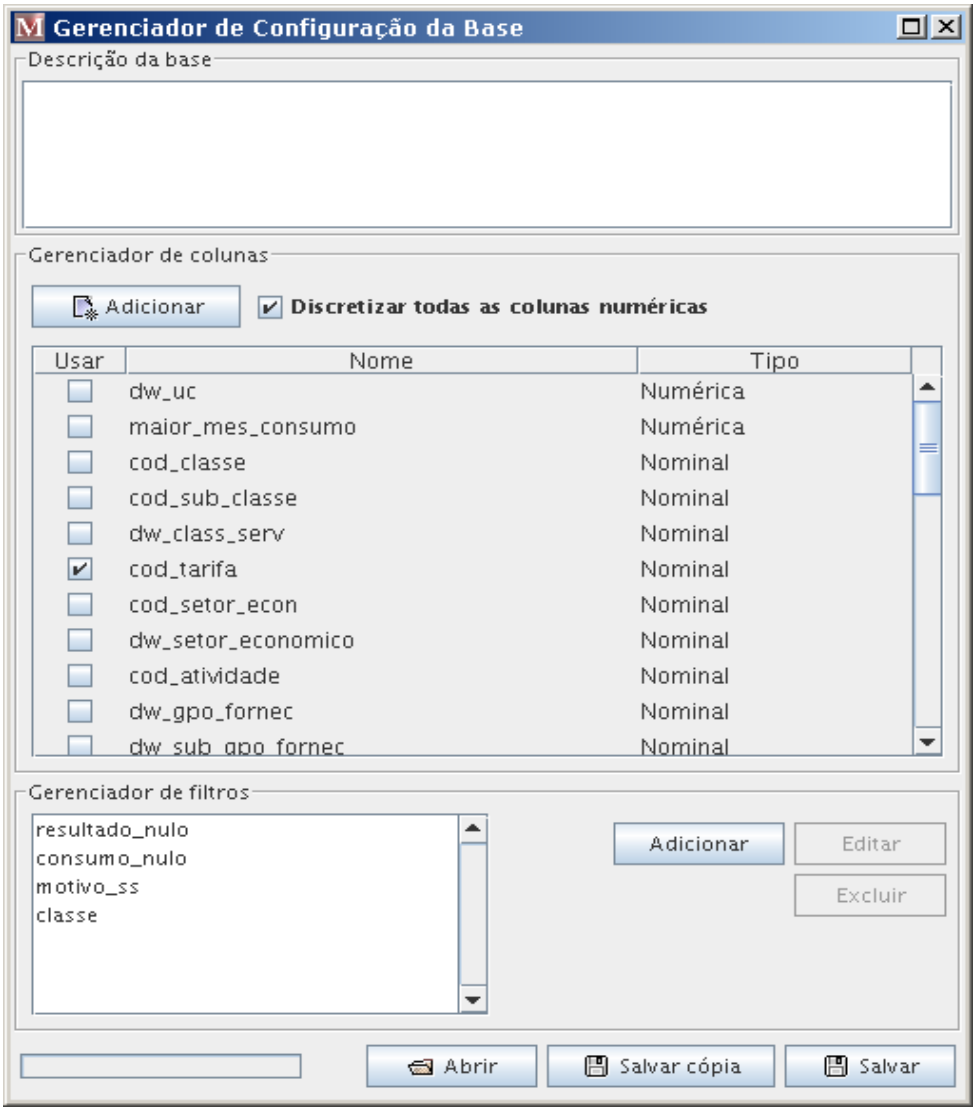

Figura 5.6: Interface de configuração da base de dados

5.4.1 Inserindo uma descrição para base

Na figura 5.7 é ilustrado o espaço da tela de gerenciamento de configurações da base destinado à uma descrição livre sobre os dados. Esta descrição será incorporada às informações dos classificadores criados no sistema, para futura identificação dos dados geradores de cada classificador. No Módulo de Aprendizagem, é possível inserir uma descrição diferente para cada tarefa criada na lista de tarefas. Para isto, basta que o usuário altere a descrição da base antes de adicionar cada tarefa.

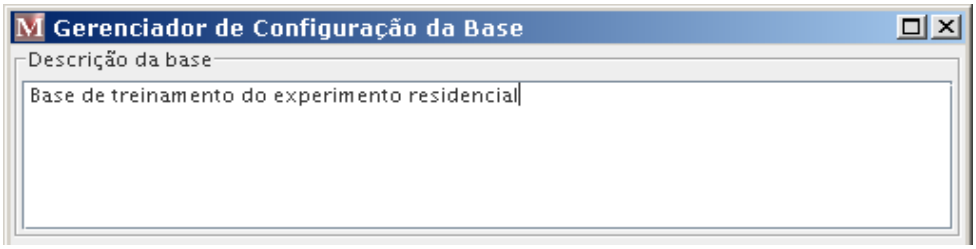

Figura 5.7: Inserindo uma descrição para a base de dados corrente

5.4.2 Gerenciando as colunas da base

Na parte central da tela de gerência da base de dados é exibido o gerenciador de colunas. Neste painel é possível editar as propriedades das colunas existentes na base de dados, adicionar novas colunas que, por não existirem no arquivo de entrada do sistema, deverão ser calculadas a partir das colunas existentes, e gerenciar o modo de tratamento dos dados numéricos da base de dados.

Na figura 5.8 é exibida a parte do gerenciador de colunas no qual é possível editar as propriedades das colunas existentes na base de dados. A seguir são explicadas quais as propriedades das colunas podem ser alteradas.

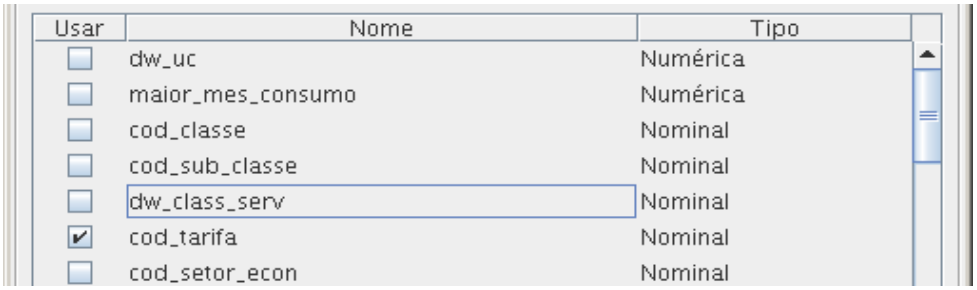

Figura 5.8: Gerenciando as colunas da base de dados

Alterando o tipo de uma coluna

Uma coluna que já exista na base de dados de entrada no sistema pode ser de dois tipos: numérica ou nominal. Uma coluna numérica armazenará um número. Uma coluna nominal armazenará um caractere (que eventualmente também pode ser um número) que representa um valor simbólico do exemplo em questão.

Por exemplo, supondo que a base de dados de entrada do sistema contenha duas colunas: altura e cor do cabelo. A coluna altura será uma coluna numérica, pois conterá um número real que indica quantos metros o indivíduo representado pelo exemplo atual tem. É possível ordenar os exemplos por altura, por exemplo.

Já a coluna cor do cabelo terá valores como “preto”, “vermelho”, “castanho” etc. Eventualmente, a coluna cor do cabelo poderia conter códigos que representassem as cores (por exemplo 1 para preto, 2 para vermelho etc.), porém, mesmo neste caso, não é correto considerar a coluna como numérica, dado que não existe a idéia de ordem e distância neste conceito.

Na figura 5.9 é mostrado como é possível alterar o tipo das colunas da base de dados do sistema.

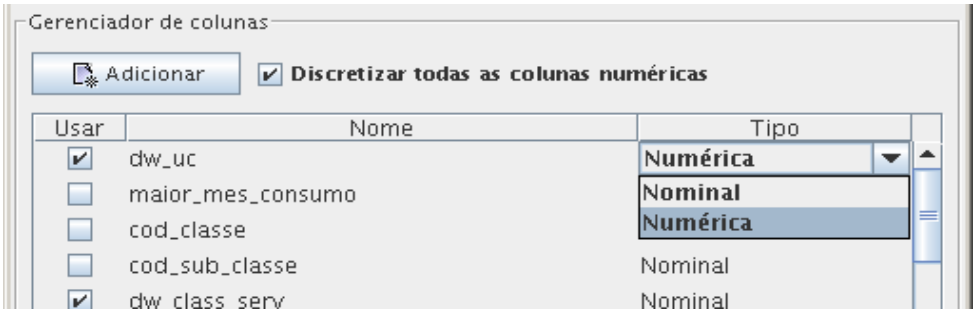

Figura 5.9: Alterando o tipo de uma coluna

Como exceção, o sistema não permite que o tipo da última coluna do arquivo de entrada seja alterado. Esta coluna necessariamente deve ser do tipo nominal.

Escolhendo entre dados numéricos discretizados ou não discretizados

Os dados das colunas numéricas podem também ser convertidos para um formato simbólico através de um processo de discretização. Um exemplo de processo de discretização é a estratificação de resultados de pesquisas demográficas dos institutos oficiais. Por exemplo, costuma-se divulgar o resultado da renda da população por faixas: indivíduos que recebem até um limite estão na faixa 1; indivíduos que recebem acima deste limite e abaixo de um segundo limite estão da faixa 2 etc.

Na figura 5.10 é ilustrado como instruir o sistema a converter colunas numéricas em colunas simbólicas (por processo de discretização). O sistema criará automaticamente sete faixas para cada coluna numérica, contendo exatamente a mesma quantidade de exemplos em cada faixa. Marque a caixa da figura 5.10 se você deseja discretizar as colunas numéricas. É importante ressaltar que nenhuma ação será realizada sobre as colunas configuradas como sendo do tipo nominal.

Figura 5.10: Selecionando entre dados numéricos discretizados ou não

Experimentos com os dados dos consumidores da ESCELSA indicam que algumas técnicas de classificação obtém melhores resultados quando trabalhando com colunas discretizadas e outras obtém melhores resultadas com as colunas no formato original.

As seguintes técnicas apresentaram melhores resultados experimentais com a base de dados em formato numérico (caixa da figura 5.10 desmarcada):

- Comparador de curvas
- J48
- Knn

As demais técnicas apresentaram melhores resultados experimentais com a base de dados discretizada (caixa da figura 5.10 marcada).

Caso esteja sendo montado um treinamento de vários classificadores no Módulo de Aprendizagem, cada vez que uma nova tarefa for adicionada à lista de tarefas pendentes para treinamento, uma cópia desta configuração é salva juntamente com a tarefa. Assim, é possível adicionar tarefas de treinamento contendo os dois tipos de configuração, bastando que a configuração seja alterada antes que acrescida à lista de execução.

O Módulo de Consultas ignora esta configuração, conforme descrito na seção 5.4.5.

Selecionando uma coluna para participação no treinamento

Eventualmente, uma coluna que está no arquivo de entrada do sistema pode não ser mais necessária ao treinamento dos classificadores (por exemplo, porque descobriu-se experimentalmente que a coluna não melhora o desempenho dos classificadores). Neste caso, é possível desmarcar a

coluna para que ela não seja usada no treinamento. A vantagem desta opção é que o arquivo de entrada do sistema não precisará ser modificado.

Na figura 5.11, quatro colunas são exibidas. As duas primeiras colunas não participarão do treinamento. Caso queira que estas colunas também participem do treinamento, basta marcar a caixa de seleção à esquerda do nome da coluna.

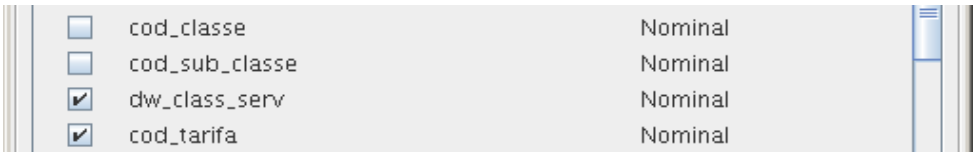

Figura 5.11: Selecionando uma coluna para participação do treinamento

Novamente, o sistema não permite que a última coluna da base - que por convenção contém a classe - seja retirada do treinamento.

Colunas temporais

O sistema MIP está preparado para manipular séries temporais. Séries temporais são, na verdade, um conjunto de valores que representam o mesmo conceito, porém em momentos de tempo distintos. Por exemplo, a medida do consumo de energia de um consumidor é uma série temporal.

Uma série temporal será representada na lista de colunas pelo nome seguido da quantidade de valores que compõe a série. Um exemplo é ilustrado na figura 5.12. Neste exemplo, o sistema identificou duas séries temporais: uma chamada “calculo” e outra chamada “curva”. Ambas são compostas por doze valores.

O sistema considera uma série temporal sempre composta de valores numéricos e não permite que o tipo da série temporal seja alterado.

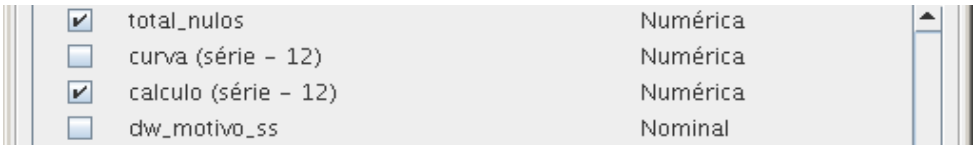

Figura 5.12: O sistema identificou duas séries temporais nos arquivos

O sistema identificará que um conjunto de colunas é uma série temporal no arquivo de entrada, quando várias colunas consecutivas possuírem o mesmo nome, exceto por um valor numérico após o nome da coluna. Por exemplo:

```
dw_uc,dw_tarifa,consumo1,consumo2,consumo3,resultado
1,2,100,120,120,1
3,2,120,180,130,1
```

No arquivo acima, o sistema identificaria que existe uma série temporal chamada “consumo” de tamanho igual a três.

Colunas calculadas

Além das colunas inicialmente disponíveis no arquivo de entrada do sistema, é possível adicionar novas colunas à base, através de funções que gerarão novos valores a partir das colunas

inicialmente disponíveis.

As novas colunas também aparecerão na lista do painel do gerenciamento de colunas, conforme ilustrado na figura 5.13.

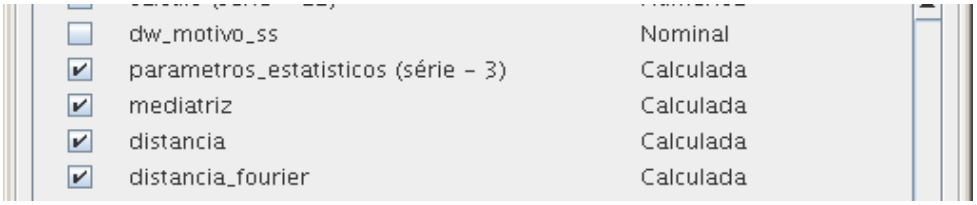

Figura 5.13: Colunas calculadas na lista de colunas

Uma coluna calculada é identificada na lista de colunas através da palavra “calculada” no campo “tipo”. Ao contrário das colunas inicialmente disponíveis no arquivo, as colunas calculadas não podem ter o tipo alterado. O tipo destas colunas (nominal ou numérico) é definido automaticamente pela função que gerará a coluna.

Conforme mostrado a seguir, para criar uma nova coluna calculada é necessário informar a função de cálculo que a gerará e as colunas inicialmente existentes que serão usadas como parâmetros para geração do novo valor.

Após criada, é possível obter as informações dos parâmetros da coluna parando o mouse sobre seu o nome na tabela do gerenciador. Por exemplo, na figura 5.14 são mostrados os detalhes da coluna “distancia_fourier”. Esta coluna foi calculada usando a função de cálculo “Distância de Fourier” e recebeu como parâmetros duas colunas que representam séries temporais: “calculo” e “curva”.

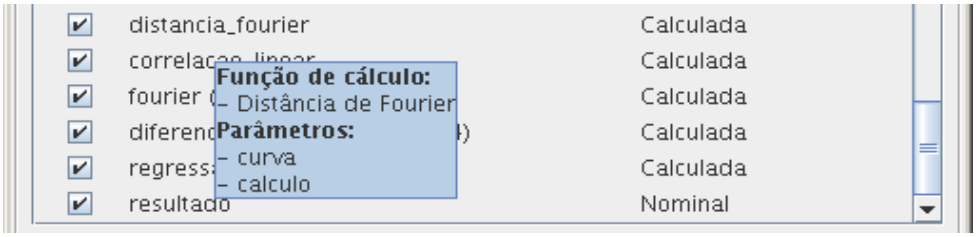

Figura 5.14: Informações sobre uma coluna calculada

Nas seções seguintes serão mostradas todas as funções de cálculo disponíveis.

Excluindo uma coluna calculada

Conforme já mostrado, uma coluna que já existe no arquivo de entrada do sistema pode ser ignorada no momento do treinamento. Para isto basta desmarcar a coluna no gerenciador.

Uma coluna calculada não pode existir na configuração e mesmo assim não ser utilizada no treinamento dos classificadores. Assim, sempre que se tentar desmarcar uma coluna calculada, o sistema informará que a coluna, na verdade, será removida da configuração, conforme ilustrado na figura 5.15.

Caso o usuário responda “Sim” na tela da figura 5.15, a coluna calculada será removida da configuração.

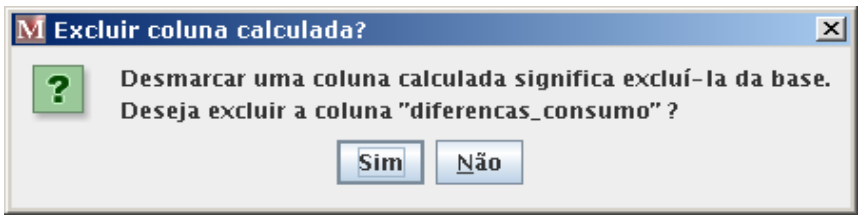

Figura 5.15: Excluindo uma coluna calculada

Inserindo uma nova coluna calcula

Para inserir uma nova coluna calculada na configuração da base atual basta clicar no botão "Adicionar" do gerenciador de colunas. Será, então, mostrada a tela ilustrada na figura 5.16.

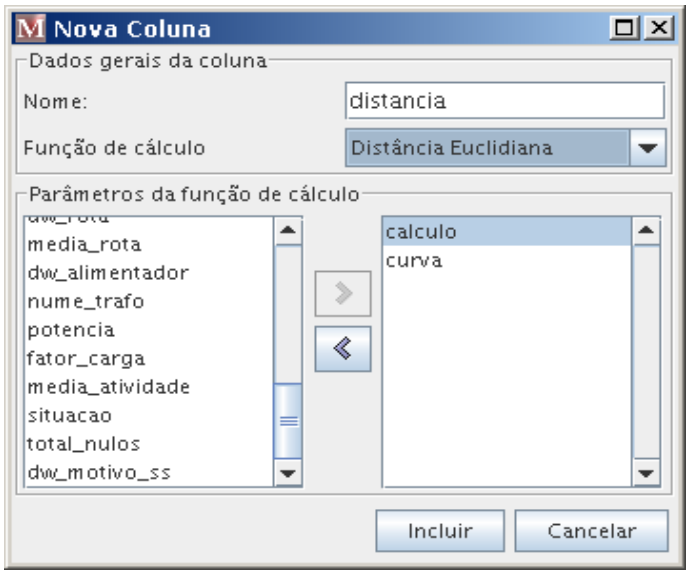

Figura 5.16: Inserindo uma nova coluna calculada

Nesta tela é necessário informar o nome da nova coluna (que não pode ser o mesmo de uma outra coluna já existente na base - calculada ou não), a função de cálculo da coluna e os parâmetros da função de cálculo. Para adicionar um parâmetro basta selecionar o nome de uma coluna na lista do lado esquerdo da tela e então clicar no botão "seta à direita". Para remover um parâmetro já adicionado, basta clicar no nome da coluna na listagem do lado direito da tela e então clicar no botão "seta à esquerda".

É obrigatório selecionar uma função de cálculo e pelo menos um parâmetro. Cada função de cálculo produz uma coluna com valor diferente e requer um conjunto de parâmetros especiais. Na lista a seguir são descritas todas as funções de cálculo disponíveis e os parâmetros que cada uma delas requer.

1. Mediatriz: esta função deve receber como entrada uma lista de colunas (pelo menos uma coluna), que podem ser séries temporais ou não. Produz como saída um valor numérico que é a média entre o máximo e o mínimo valor nas colunas passadas como parâmetros.

2. Distância Euclidiana: esta função deve receber como entrada exatamente duas séries temporais do mesmo tamanho. Como saída, produzirá um valor numérico que é a distância

média ponto a ponto dos valores que compõe a série temporal. As duas séries temporais serão normalizadas pelo maior valor de cada série, antes do cálculo da distância.

3. Amplitudes de Fourier: esta função deve receber como entrada exatamente uma série temporal. Embora não exista restrição quanto ao tamanho da série, recomenda-se não utilizar esta função com séries temporais de tamanho inferior a dez. Esta função produz como saída outra série temporal composta pelas amplitudes dos harmônicos calculados pela Transformada de Fourier.

4. Distância de Fourier: esta função deve receber como entrada exatamente duas séries temporais do mesmo tamanho. Como saída, produzirá um valor numérico que é a distância média ponto a ponto das amplitudes da Transformada de Fourier das séries. As amplitudes das Transformadas de Fourier são normalizadas pelo valor da amplitude do primeiro harmônico, antes do cálculo da distância.

5. Regressão linear: esta função deve receber como entrada exatamente uma série temporal. Produzirá como saída dois valores numéricos, que são os coeficientes do polinômio de primeiro grau obtido com a aproximação da série pelo método dos quadrados mínimos.

6. Correlação linear: esta função deve receber como entrada exatamente duas séries temporais. Produzirá como saída um valor numérico que é a correlação linear das duas séries. O cálculo da correlação linear é feito supondo que cada série é uma variável aleatória. A correlação será um valor que varia entre -1 e 1: valores próximos de -1 representam que as séries têm comportamento linear inverso (têm tendências inversas); valores próximos a 1 significam que as séries têm comportamento linear direto (têm a mesma tendência); valores próximos a 0 significam que as séries não têm comportamento linear semelhante.

7. Parâmetros estatísticos: esta função deve receber como parâmetros de entrada uma ou mais colunas (inclusive séries temporais). Produzirá como saída três valores: a média dos valores das colunas entradas, o desvio padrão dos valores de entrada e o desvio padrão normalizado dos valores de entrada. O desvio padrão normalizado é o desvio padrão dividido pelo maior valor existente nos parâmetros de entrada da função.

8. Diferenças em séries temporais: esta função deve receber como entrada exatamente uma série temporal. Produzirá como saída quatro valores: dois valores representando as maiores diferenças entre dois valores consecutivos da série temporal (um com a maior diferença absoluta e outro com a maior diferença percentual) e dois valores representando as menores diferenças entre dois valores consecutivos da série temporal (novamente, um representando a menor diferença absoluta e outro representando a menor diferença percentual).

Escolhido um nome, uma função de cálculo e os parâmetros da função de cálculo, é possível adicionar a coluna calculada clicando no botão “Incluir” da tela da figura 5.16. A nova coluna será inserida na penúltima posição da tabela do gerenciador de colunas, imediatamente antes da coluna que contém a classe dos exemplos.

5.4.3 Gerenciando os filtros de descarte

A terceira parte da gerência de configuração da base de dados e o gerenciador de filtros de descarte. Filtros de Descarte é uma coleção de filtros que será aplicada aos exemplos da base de entrada do sistema para decidir se o exemplo deve ou não participar do treinamento.

Enquanto no gerenciador de colunas era possível selecionar uma coluna da base para participação do treinamento, aqui é possível selecionar função que decidirá se uma linha da base de dados deverá participar do treinamento.

Na figura 5.17 é ilustrada a lista de filtros de descarte criada para aplicação na base corrente. Durante a execução do treinamento e da consulta, cada exemplo será testado por cada um dos filtros cadastrados. Se algum dos filtros decidir que o exemplo deve ser eliminado da base, então aquele exemplo não mais participará do treinamento.

Figura 5.17: Lista de filtros de descartes

Assim como as colunas calculadas, um filtro de descarte é composto por uma função de descarte e por parâmetros de cálculo. No caso dos filtros de descarte, é possível que valores constantes sejam também passados ao filtro, além das colunas existentes na base.

Para obter informações sobre um filtro de descarte cadastrado, pare o mouse sobre o nome do filtro na lista da tela de gerência de configurações da base. Na figura 5.18 são exibidas as informações do filtro “resultado_nulo”. Este filtro é decidido pela função “Descarte de Nulos” e recebe como parâmetro a coluna “dw_motivo_ss”. Nenhuma constante é passada como parâmetro para este filtro.

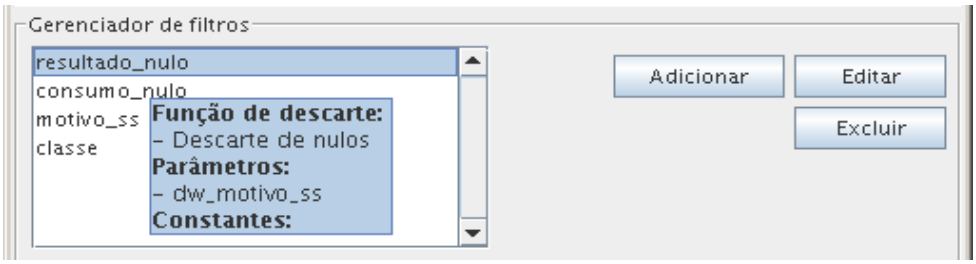

Figura 5.18: Informações sobre um filtro de descarte

Excluindo um filtro de descarte

Para excluir um filtro de descarte da lista de filtros, basta clicar no nome do filtro que o botão “Excluir” do painel do gerenciador de filtros será habilitado. O botão “Excluir” está ilustrado na figura 5.19.

Inserindo um novo filtro de descarte

Para inserir um novo filtro de descarte, basta clicar no botão “Adicionar” do painel do gerenciador de filtros (veja figura 5.19). Será exibida a tela ilustrada na figura 5.20.

Para inserir um novo filtro de descarte é necessário: digitar um nome para o filtro (que não pode ser igual ao nome de outro filtro já existente), selecionar uma função de descarte, digitar

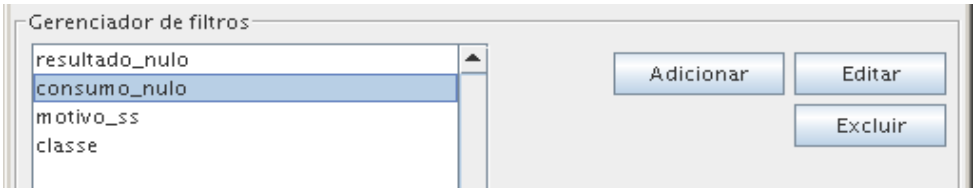

Figura 5.19: Excluindo um filtro de descarte

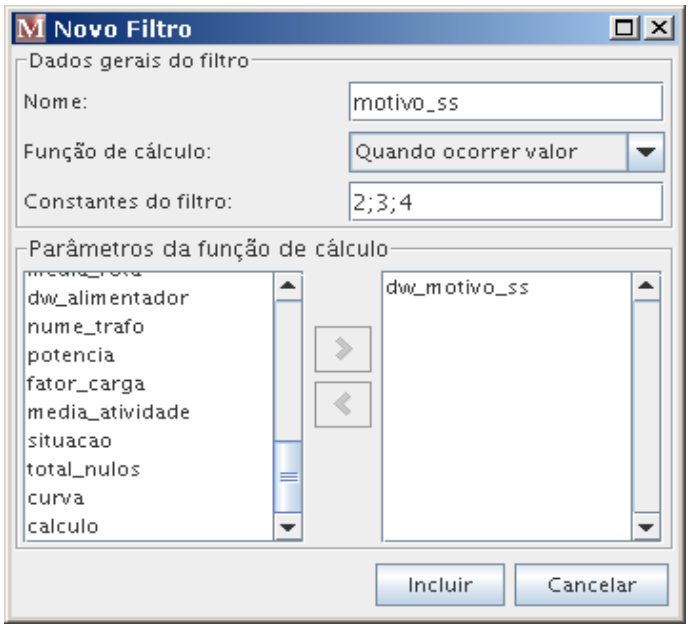

Figura 5.20: Inserindo um novo filtro de descarte

as constantes a serem usadas como parâmetro do filtro e selecionar as colunas da base que serão usadas como parâmetro do filtro.

Os parâmetros constantes do filtro devem ser digitados como uma lista de valores separados por ponto e vírgula. Na figura 5.20, o novo filtro cadastrado recebe como parâmetro três valores: 2;3;4. Notar que os valores estão separados por ponto e vírgula.

Cada função de descarte requer parâmetros especiais (exatamente como no caso das funções de cálculo das colunas calculadas). A seguir são listadas todas as função de descarte disponíveis e o tipo de parâmetro que elas esperam receber.

1. Nulos em série de consumo: esta função espera como parâmetro exatamente uma série temporal de tamanho maior que quatro. Esta função não deve receber parâmetros constantes. Caso seja criado um filtro usando esta função, serão descartados todos os exemplos cujo algum dos três últimos valores da série temporal passada como parâmetro seja nulo. Esta regra de descarte foi implementada especialmente para o caso do tratamento dos nulos nas séries de consumo do problema de classificação da ESCELSA, por isto a função recebeu este nome.

2. Descarte de nulos: esta função espera receber uma ou mais colunas como parâmetro (porém não deve receber nenhuma constante). As colunas passadas como parâmetro podem ser de qualquer tipo, inclusive séries temporais. Ao criar um filtro usando esta função, todos os exemplos da base que tiverem qualquer valor nulo nas colunas passadas como parâmetro

serão descartados.

3. Quando ocorrer valor: este filtro espera receber exatamente uma coluna (que não pode ser uma série temporal) e uma lista de constantes separadas por ponto e vírgula. Serão descartados todos os exemplos em que uma das constantes passadas como parâmetro ocorrer na coluna selecionada. No exemplo da figura 5.20, serão descartados todos os exemplos que possuírem dw_motivo_ss igual a 2, 3 ou 4.

4. Quando não ocorrer valor: este filtro espera exatamente os mesmos parâmetros do filtro anterior e descarta todos os exemplos em que não ocorrer alguma das constantes na coluna selecionada para ser parâmetro do filtro.

Após cadastrar o nome do filtro, selecionar a função de descarte e preencher os parâmetros do filtro (coluna e constantes), basta acionar o botão “Incluir” que o filtro será adicionado ao final da lista de filtros do painel da tela de gerenciamento da base de dados.

Editando um filtro de descarte

A edição de um filtro de descarte é feita de maneira semelhante ao cadastro de um novo filtro. Para editar um filtro de descarte, basta selecioná-lo na tela principal do sistema e acionar o botão “Editar” do gerenciador de filtros. Será aberta a tela ilustrada na figura 5.21. O funcionamento desta tela é exatamente o mesmo da tela de cadastro de novos filtros.

Após efetuar as alterações no filtro, basta clicar no botão “Salvar” que as alteração são efetivadas.

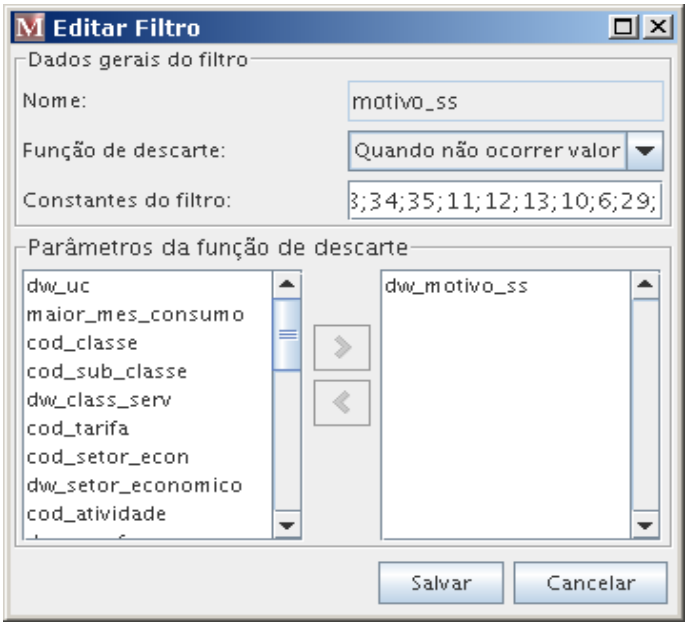

Figura 5.21: Editando um filtro de descarte

5.4.4 Salvando e recuperando configurações

Configurar uma base de dados pode ser um trabalho longo. Porém, após fazer uma configuração, espera-se que ela não seja alterada com tanta freqüência. O sistema MIP, por padrão, recarregará sempre a última configuração usada no treinamento ou na consulta.

O sistema possuiu também opções de arquivamento e recuperação de configurações. Estas opções são extremamente úteis quando o usuário necessita alternar entre formato de bases: ora quer realizar um treinamento ou uma consulta com um formato, ora quer realizar um treinamento ou uma consulta com outro formato.

Na figura 5.22 é ilustrada a barra de botões da janela de gerência da base de dados, na qual é possível salvar e recuperar configurações.

Figura 5.22: Botões de recuperação e gravação da configuração da base

Na barra de botões da figura 5.22, o botão “Salvar” efetiva todas as alterações feitas na configuração atual e fecha a janela do configurador da base de dados, retornando para o módulo que o acionou. As alterações feitas na configuração atual somente serão efetivamente usadas no sistema se o botão “Salvar” for acionado.

Quando o botão “Salvar” é acionado, o sistema checa se a configuração editada é válida para a base de dados atualmente selecionada no sistema. Caso a base de dados selecionada possua muitos exemplos, esta verificação pode demorar alguns segundos.

Ainda na barra de botões da figura 5.22, existe um botão chamado “Salvar cópia”. Este botão possibilita que a configuração atual seja salva em disco em arquivo de formato XML. Esta opção não efetiva a configuração para utilização no sistema (como faz o botão “Salvar”), apenas faz uma cópia da configuração para utilização futura.

Uma configuração salva pela opção “Salvar cópia” pode ser recuperada com a utilização do botão “Abrir”. Na figura 5.23 é ilustrada a tela de seleção de configurações salvas, acionada pelo botão “Abrir”. Após abrir uma configuração salva, é necessário acionar o botão “Salvar” para que a configuração seja efetivamente usada no sistema. Uma configuração aberta pode ser, eventualmente, alterada antes de ser usada no sistema.

A instalação do sistema MIP possui quatro configurações pré-definidas para utilização do usuário. As configurações estão disponíveis no sub-diretório “conf/modelos” do diretório principal do sistema.

As quatro configuração (cujos os arquivos são ilustrados também na figura 5.23) supõem que a base de dados estará no formato de saída do módulo de conversão de dados do Datawarehouse (veja apêndice B). Além disso, elas possuem as seguintes características:

- residencial_treinamento.xml - este arquivo contém a configuração usada para treinamento dos classificadores de seleção de consumidores residenciais da ESCELSA para inspeção. Dentre as colunas disponíveis no arquivo de entrada (cujo formato esperado é descrito no apêndice B) esta configuração seleciona para participação do treinamento as colunas cod_tarifa (tarifa do exemplo em questão), media_rota (média de consumo da rota de leitura do exemplo), potência (do transformador), fator_carga (relação entre potência e consumo do transformador), total_nulos (total de meses sem medição de consumo na série temporal), calculo (série temporal de consumo) e resultado de inspeção.

  Além disso, todas as funções de cálculo descritas são usadas para cálculo de novas colunas. Todas as novas colunas são calculas a partir da série de consumo. As funções que necessi-

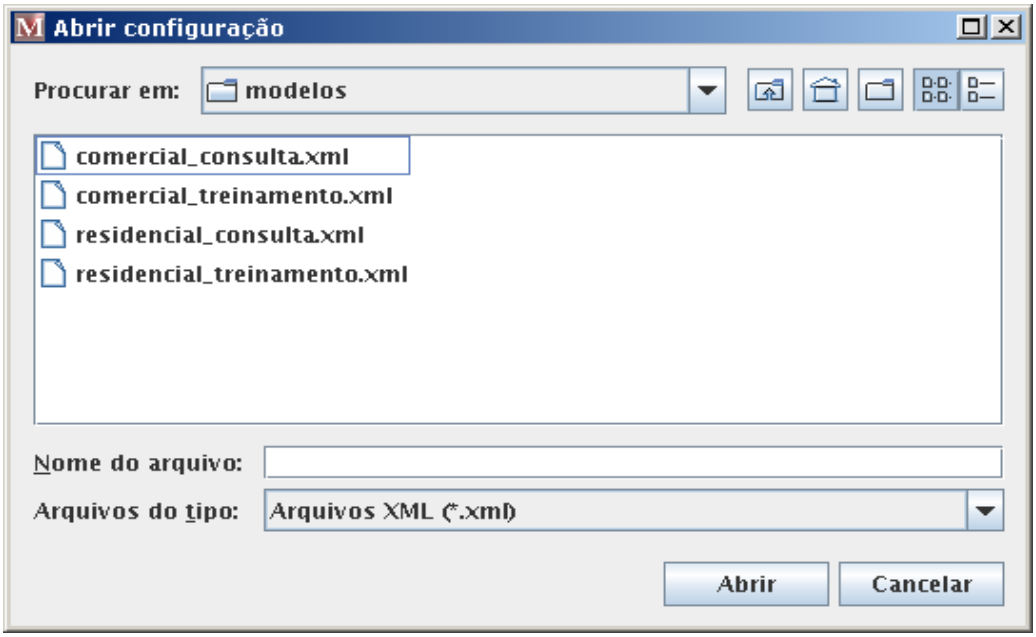

Figura 5.23: Modelos de configurações pré-definidos

tam exatamente de duas séries temporais como parâmetros recebem ainda a série de curvas típicas.

Esta configuração possui quatro filtros de descartes:

"resultado_nulo" - remove da base de exemplos consumidores que não foram inspecionados. Para isto, filtra todos os consumidores que possuírem a coluna dw_motivo_ss nula.

"consumo_nulo" - usa a função de filtragem de séries temporais para remover exemplos com nulos no final da série temporal de consumo.

"motivo_ss" - remove da base de dados exemplos que não possuem o motivo de inspeção igual a um dos valores considerados corretos pelo especialista da ESCELSA. Este filtro tem como objetivo remover ruídos da base de dados de treinamento.

"classe" - esta configuração foi feita especialmente para treinamento de classificadores para seleção de consumidores residenciais. Este filtro exclui da base de dados todos os exemplo cuja a classe não seja residencial.

- residencial_consulta.xml - este arquivo de configuração foi feito para utilização na consulta com classificadores treinados com a configuração anterior. As colunas desta configuração são as mesmas da configuração anterior. A diferença em relação à configuração anterior está nos filtros de descarte. Esta configuração possui apenas dois filtros: "consumo_nulo" e "classe". Os demais filtros não fazem sentido na consulta, dado que o resultado da inspeção ainda não é conhecido para os exemplos da base de consulta.

- comercial_treinamento.xml - este arquivo contém a configuração usada para treinamento dos classificadores de seleção de consumidores comerciais e industriais da ESCELSA para inspeção. Dentre as colunas disponíveis no arquivo de entrada (cujo formato esperado é descrito no apêndice B) esta configuração seleciona para participação do treinamento as colunas cod_tarifa (tarifa do exemplo em questão), cod_setor_econ (setor econômico do exemplo), cod_atividade (código da atividade do consumidor), potência (do transformador), fator_carga

(relação entre potência e consumo do transformador), media_atividade (média de consumo da atividade do exemplo), total_nulos (total de meses sem medição de consumo na série temporal), calculo (série temporal de consumo) e resultado de inspeção.

Todas as funções de cálculo também são usadas para criação de novas colunas, exatamente da mesma forma que na configuração anterior.

Esta configuração possui três filtros de descartes:

"resultado_nulo" - remove da base de exemplos consumidores que não foram inspecionados. Para isto, filtra todos os consumidores que possuírem a coluna dw_motivo_ss nula.

"consumo_nulo" - usa a função de filtragem de séries temporais para remover exemplos com nulos no final da série temporal de consumo.

"classe" - esta configuração foi feita especialmente para treinamento de classificadores para seleção de consumidores comerciais e industriais. Este filtro exclui da base de dados todos os exemplo cuja a classe não seja comercial ou industrial.

- comercial_consulta.xml - esta configuração foi feita para utilização na consulta com classificadores criados com a configuração anterior. A configuração é exatamente a mesma que a do item anterior, exceto pelos filtros de descartes. Nesta configuração o filtro "resultado_nulo" não é aplicado, dado que os exemplos não possuem ainda resultado de inspeção.

5.4.5 Considerações sobre a configuração da base para consultas

Nas seções anteriores foi descrito em detalhes o funcionamento da tela de configuração da base de dados. Conforme notado, o sistema possibilita uma grande flexibilidade para que o usuário selecione arquivos nos mais diferentes formatos e configure, a seu gosto, como será a base usada no treinamento e como será a base usada na consulta.

Além das restrições em relação ao formato CSV, alguns cuidados devem ser tomados quando estiver sendo configurada a base para utilização no Módulo de Consultas. Em primeiro lugar, o objetivo do Módulo de Consultas é descobrir a qual classe um exemplo desconhecido pertence, dado um classificador previamente treinado. Seria lógico, portanto, imaginar que a coluna que contém a classe não é necessária na base de dados de consulta.

Embora esta coluna não seja utilizada no processo de consulta, a base de dados apresentada ao Módulo de Consultas deve, obrigatoriamente, conter esta coluna. Os valores da coluna, por obviamente não serem conhecidos, devem estar marcados como desconhecidos (ver seção 5.2).

Além disto, a base de consultas deve, necessariamente, conter as mesmas colunas da base de treinamento. Isso significa que não é possível treinar um classificador com uma base que contém algumas colunas (por exemplo usando a configuração anterior para consumidores residenciais) e consultar com uma base que contenha outras colunas (por exemplo, a base anterior para consumidores comerciais e industriais).

Por fim, as configurações de discretização no Módulo de Consultas são obtidas a partir da configuração de cada classificador e não da interface de configuração da base de dados. Assim, esta configuração pode ser ignorada no Módulo de Consultas.

5.5 Selecionando uma base em formato incompatível com a configuração atual do sistema

Quando uma configuração de base de dados estiver sendo usada e o usuário selecionar uma base (tanto em formato CSV quando no formato de saída do Datawarehouse) o sistema verificará se a base de dados selecionada é compatível com a configuração atual. Caso a base não seja compatível, será exibida a mensagem informativa ilustrada na figura 5.24.

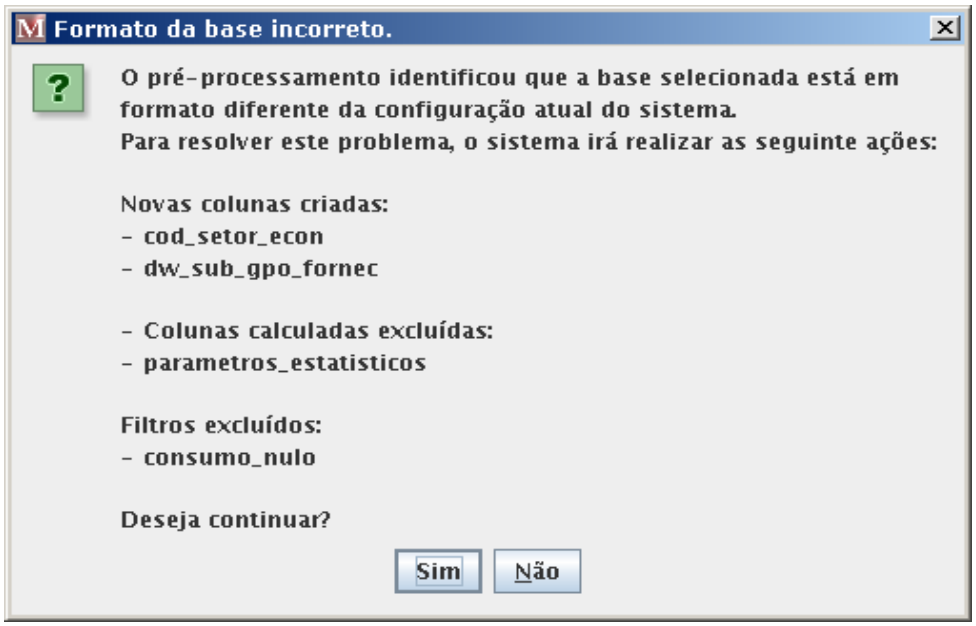
Formato da base incorreto.
O pré-processamento identificou que a base selecionada está em formato diferente da configuração atual do sistema.
Para resolver este problema, o sistema irá realizar as seguinte ações:
Novas colunas criadas:
- cod_setor_econ
- dw_sub_gpo_fornec
- Colunas calculadas excluídas:
- parametros_estatisticos
Filtros excluídos:
- consumo_nulo
Deseja continuar?
Sim
Não

Figura 5.24: Selecionando uma base incompatível com a configuração atual do sistema

Uma base é incompatível se alguma das colunas da configuração atual não existir na base que está sendo aberta (isto é checado pelo nome da coluna, que no caso do arquivo CSV, está na primeira linha do arquivo) ou se a base que está sendo aberta contém colunas novas, que não existiam na configuração anterior.

Caso o usuário clique no botão “Sim” da tela ilustrada na figura 5.24, a tela de gerência de configuração da base será aberta e o sistema tentará criar automaticamente uma configuração compatível com a nova base. Normalmente, o sistema só não conseguirá criar um configuração compatível se alguns dos problemas descritos na seção 5.2 for identificado.

5.6 Selecionando uma base de dados em formato de saída do Datawarehouse

Na figura 5.25 é mostrada a forma de acesso à interface de seleção de base de dados no formato de saída do Datawarehouse. Está sendo chamado de “formato de saída do Datawarehouse” o conjunto de sete arquivos disponibilizados pela ESCELSA, extraídos a partir do Datawarehouse da empresa.

O sistema MIP possui uma interface para conversão destes sete arquivos. Nesta interface, acessada a partir do menu da figura 5.25, é possível selecionar os arquivos gerados pelo Datawarehouse e iniciar a conversão. O produto final da conversão será uma base de dados em formato CSV, exatamente como as descritas nas seções anteriores. Na figura 5.26 é ilustrada a interface

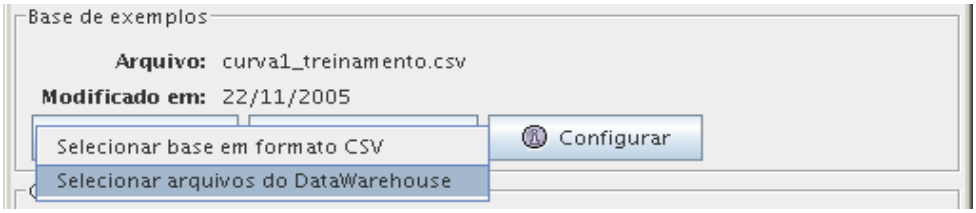

Figura 5.25: Seleção de base no formato de saída do Datawarehouse

principal da funcionalidade de conversão da base.

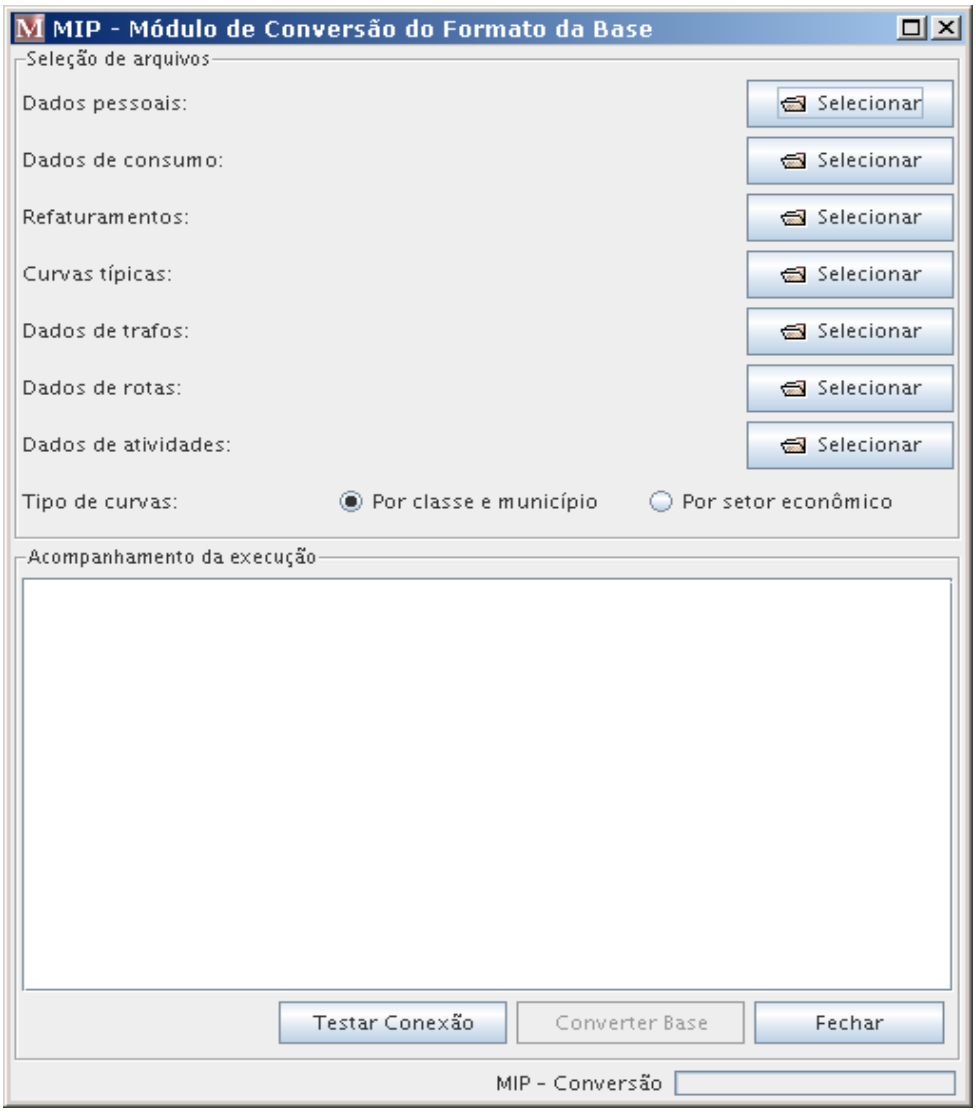

Figura 5.26: Tela principal da interface de conversão

O primeiro passo para iniciar uma conversão é selecionar os sete arquivos disponibilizados pela ESCELSA. Na figura 5.27 é ilustrado o painel da interface no qual deve ser feita a seleção dos arquivos.

Os sete arquivos a selecionar são: arquivo com dados pessoais, arquivo com dados de consumo, arquivo com dados de refaturamento, arquivo com dados de curva típicas, arquivo com dados de transformadores, arquivo com dados de rotas e arquivos com dados de atividades. O formato esperado de cada um dos arquivos é descrito no apêndice A.

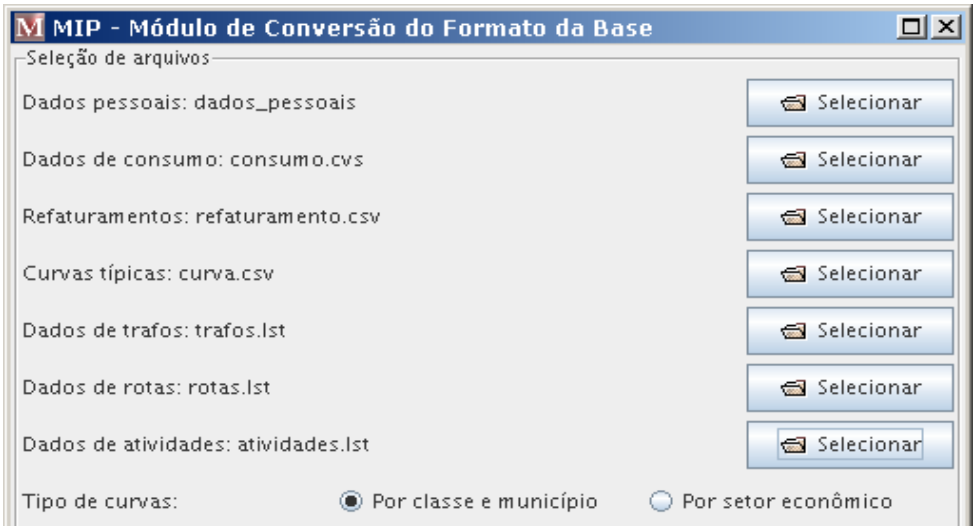

Figura 5.27: Selecionando os arquivos para conversão

Para iniciar uma conversão é necessário que o banco de dados PostgreSQL esteja instalado e em execução. Na interface de conversão existe um botão que verifica se a conexão com o banco de dados está funcionando corretamente. Para verificar se o servidor de banco de dados encontra-se em funcionamento normal, utilize o botão “Testar Conexão” da tela de conversão. Caso a conexão esteja funcionando corretamente, o sistema mostrará a mensagem ilustrada na figura 5.28.

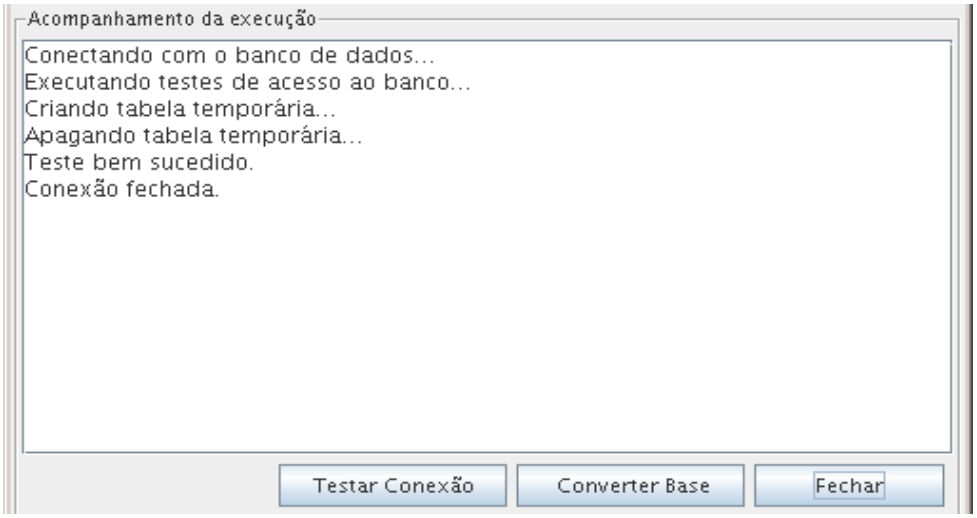

Figura 5.28: Testando a conexão com o banco de dados

Caso o sistema exiba no painel de informações a mensagem ilustrada na figura 5.29, o servidor de banco de dados não está em execução ou o sistema MIP não teve acesso ao servidor. Para resolver este problema consulte os procedimentos de instalação do servidor do banco de dados, na seção 2.

Ainda na tela de conversão, existe uma opção para seleção do tipo de curva típica a ser usada na conversão. Existem duas opções possíveis: curva típica por setor econômico ou curva típica por classe e município. Vale destacar que o arquivo de entrada a ser selecionado no campo de seleção do arquivo de curva típica, mostrado na figura 5.27, varia de acordo com a opção de curva selecionada. No apêndice A é descrito o formato dos dois tipos de arquivos possíveis.

O arquivo produzido pelo conversor terá sempre o mesmo formato, independente da seleção

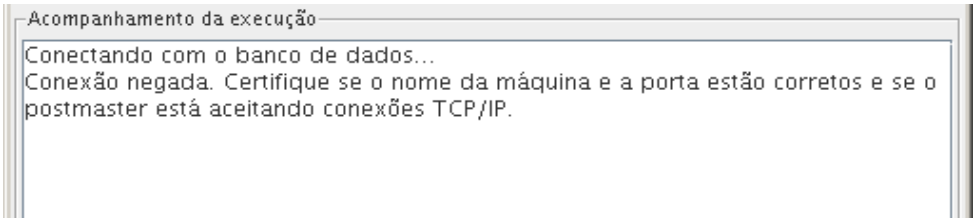
Acompanhamento da execução
Conectando com o banco de dados...
Conexão negada. Certifique se o nome da máquina e a porta estão corretos e se o postmaster está aceitando conexões TCP/IP.

Figura 5.29: Problemas com a conexão com o banco de dados

Figura 5.30: Escolhendo o tipo de curva típica

de curva típica. A principal diferença será na forma como as colunas de curva típicas serão geradas pelo sistema. No apêndice B é descrito o formato do arquivo CSV gerado, e explicada a diferença entre as duas curvas típicas existentes.

Após selecionar os sete arquivos e informar qual tipo de arquivo de curva típica está sendo usado, o usuário pode iniciar uma conversão. A conversão é iniciada com o acionamento do botão "Converter base".

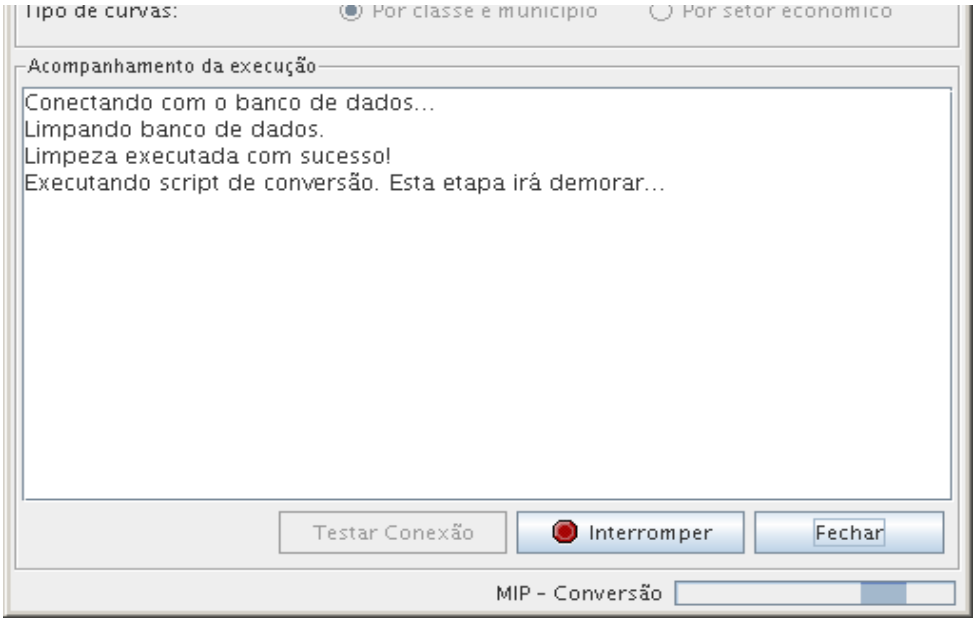

Figura 5.31: Executando uma conversão

Na figura 5.31 é mostrado o estado da tela durante a conversão de um arquivo. Normalmente, a conversão dos arquivos demora um longo tempo. Caso os arquivos sejam grandes, com mais de 20.000 casos, por exemplo, é normal que a conversão demore mais de 10 minutos para ser concluída. A qualquer momento o usuário pode interromper o processo usando o botão "Interromper" que será mostrado no local do botão "Converter base". O sistema informará na caixa de acompanhamento da execução os possíveis erros ocorridos durante a conversão.

Ao finalizar a conversão, será exibida a mensagem ilustrada na figura 5.32. Caso a mensagem seja respondida com a opção "Sim", o sistema retornará para a interface principal do módulo que acionou a seleção de dados do Datawarehouse já com o arquivo selecionado para utilização. Se,

eventualmente, for necessário configurar a base (por exemplo, se a base for incompatível com a última configuração salva no sistema), os procedimentos deste ponto em diante são exatamente os mesmos dos descritos para manipulação do arquivo de entrada em formato CSV.

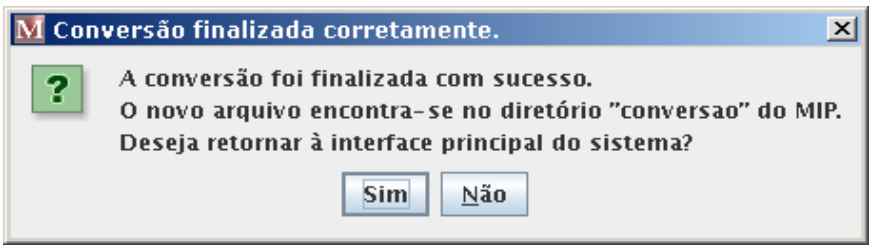
Conversão finalizada corretamente.
A conversão foi finalizada com sucesso.
O novo arquivo encontra-se no diretório "conversao" do MIP.
Deseja retornar à interface principal do sistema?
Sim
Não

Figura 5.32: Abrindo o arquivo gerado na interface principal do sistema

Caso o usuário queira realizar outra conversão, basta responder “Não” na mensagem da figura 5.32 e continuar usando a interface de conversão normalmente. A conversão anterior estará salva no diretório “conversao” do sistema MIP.

Apêndice A

Formato dos arquivos de entrada do conversor

O conversor dos arquivos de saída do Datawarehouse espera como entrada sete arquivos: arquivo com dados pessoais, arquivo com dados de consumo, arquivo com dados de refaturamento, arquivo com dados de curva típicas, arquivo com dados de transformadores, arquivo com dados de rotas e arquivos com dados de atividades. Nesta seção é descrito o formato esperado destes arquivos. O formato aqui descrito é o formato da instalação do sistema MIP. Porém é possível reprogramar o procedimento de conversão conforme descrito na seção C. É importante ressaltar que qualquer alteração no formato dos arquivo aqui descritos, sem os devidos acertos no processo de conversão, fará com que o conversor descrito na seção 5 pare de funcionar.

Todos os sete arquivos deverão estar em formato texto. Cada linha deverá conter a série de valores separados por ponto e vírgula. Não devem haver cabeçalhos com o nome das colunas. Os arquivos de saída do Datawarehouse devem possuir apenas dados, desde a primeira linha de seu conteúdo.

Caso algum valor de um registro esteja ausente (nulo) a posição do valor deve estar vazia, porém os separadores (ponto e vírgula) deverão aparecer normalmente. Por exemplo, os registros abaixo são válidos e possuem nulos na segunda posição (primeira linha) e na terceira (última) posição (segunda linha):

```
288;;1
2260;3;
```

A seguir o formato de cada arquivo é brevemente descrito. A ordem em que as colunas são apresentadas é a ordem na qual se espera que elas estejam nos arquivos de dados.

1. Arquivo com dados pessoais - este arquivo contém a lista de consumidores que se deseja usar no sistema (treinamento ou consulta). Cada consumidor deve estar em uma linha do arquivo. O arquivo é composto pelas seguintes colunas:

    - dw_uc - chave de identificação única dos consumidores no banco de dados da ESCELSA
    - nro_mes_oco - esta coluna não possui nenhum valor útil. Ela ainda consta no arquivo apenas para manutenção de compatibilidade com versões anteriores do sistema.
    - dat_exe - data da execução da última inspeção na residência do cliente
    - dw_cate_cliente - chave de identificação da categoria do cliente

- dw_tarifa - chave de identificação da tarifa do cliente
- cod_tarifa - código da tarifa do cliente
- dw_classe - chave de identificação da classe a qual o cliente pertence (residencial, comercial etc.)
- cod_classe - código da classe a qual o cliente pertence
- dw_sub_classe - chave de identificação da sub-classe a qual o cliente pertence
- cod_sub_classe - código da sub-classe a qual o cliente pertence
- dw_class_serv - chave de identificação da classe de serviço a qual o cliente pertence (monofásico, bifásico etc.)
- dw_setor_economico - chave de identificação do setor econômico do cliente
- cod_setor_econ - código do setor econômico do cliente
- dw_sub_classe_aneel - chave de identificação da sub-classe do cliente segundo classificação da ANEEL
- dw_atividade - chave de identificação da atividade do cliente
- cod_atividade - código da atividade do cliente
- dw_result_inspecao - resultado da última inspeção na residência do cliente
- dw_pt - chave de identificação do posto do cliente
- nume_trafo - número de identificação do transformador que atende o cliente
- feat_num_pt - coluna não utilizada. Também é mantida por questões de compatibilidade
- dsc_pt - coluna não utilizada. Também é mantida por questões de compatibilidade
- feat_num_ps - coluna não utilizada. Também é mantida por questões de compatibilidade
- dsc_ps - coluna não utilizada. Também é mantida por questões de compatibilidade
- dw_motivo_ss - chave de identificação do motivo da última inspeção na residência do cliente
- dw_sub_gpo_fornec - chave de identificação do sub-grupo de fornecimento do cliente
- dat_lig - data de ligação da unidade consumidora do cliente
- num_documento - coluna não utilizada. Também é mantida por questões de compatibilidade
- dw_rota - chave da identificação da rota de leitura a qual pertence o cliente
- dw_bairro_pt - chave de identificação do bairro da unidade consumidora do cliente
- nome_bairro - nome do bairro
- dw_alimentador - chave de identificação do alimentador que fornece energia ao cliente
- dw_pot_nomi_eqp - chave de identificação da potência nominal do transformador ao qual está ligado o cliente
- pot_nomi - potência nominal do transformador que atende o cliente
- dw_municipio_com - chave de identificação do município da unidade consumidora do cliente

- dsc_municipio - descrição do município
- dw_gpo_fornec - chave de identificação do grupo de fornecimento do cliente
- cod_sub_gpo_fornec - código do grupo de fornecimento
- dw_area_atividade - chave de identificação da área de atividade do cliente
- cod_area_ativid - código da área de atividade do cliente
- dw_tipo_grupo - tipo de grupo do cliente
- dw_situacao_uc - situação atual da unidade consumidora do cliente (ligada / desligada)
- dw_classific_uc - classificação da unidade consumidora
- dw_tipo_usuario_sistema - tipo do usuário do sistema que gerou a inspeção (funcionário, terceiro etc.)

Este arquivo não deve ter ponto e vírgula no final de cada registro.

2. Arquivo com dados de consumo - este arquivo contém os dados do consumo de energia, de irregularidades comerciais e de ocorrência de faturamento dos consumidores. Neste arquivo, cada linha conterá os registros de um consumidor em um determinado mês/ano. O arquivo será composto das seguintes colunas:

   - dw_uc - chave de identificação do consumidor
   - nro_mes_oco - mês de ocorrência do consumo, irregularidade e refaturamento do consumidor
   - dw_irreg_leitura - código da irregularidade no mês
   - dw_ocorrencia_fat - código da ocorrência de refaturamento no mês
   - calc_energia - valor do cálculo de energia no mês

   No momento da criação do processo de extração deste arquivo, foi convencionado que cada linha do arquivo conterá um ponto e vírgula no final. Assim, as duas linhas abaixo são exemplos de registros válidos para este arquivo.

   ```
   60855;200201;0;0;371;
   60855;200202;0;0;323;
   ```

3. Arquivo com dados de refaturamento - este arquivo contém a relação de ajustes nas faturas dos consumidores, registrados no banco de dados da ESCELSA. Esta informação serve como correção da informação de consumo. Caso algum consumidor tenha recebido a fatura com algum problema (crédito ou débito indevido), a correção aparecerá neste arquivo. O arquivo é composto pelas seguintes colunas:

   - dw_uc - chave de identificação do consumidor
   - nro_mes_oco - mês de ocorrência do refaturamento
   - saldo - total de consumo que deve ser reajustado na fatura original
   - dw_tipo_fatura - tipo de ajuste (crédito ou débito)
   - desc_tipo_fatura - descrição do tipo de ajuste

No momento da criação do processo de extração deste arquivo, foi convencionado que cada linha do arquivo conterá um ponto e vírgula no final. Assim, as duas linhas abaixo são exemplos de registros válidos para este arquivo.

```
530592;200502;79;1;DEBITO VALIDO;
530592;200502;79;2;DEBITO ESTORNADO;
```

4. Arquivo com dados de transformadores - este arquivo contém uma relação de todos os transformadores existentes no banco de dados da ESCELSA, com algumas informações relativas aos transformadores. O arquivo deve conter as seguintes colunas:

   - dw_trafo - chave primária do transformador no banco de dados da ESCELSA.
   - dw_pt - chave primária do posto do transformador no banco de dados da ESCELSA.
   - total_consumo - soma do consumo de todos os clientes do transformador no último mês.
   - potencia_trafo - potência nominal do transformador

   Este arquivo não deve ter ponto e vírgula no final de cada registro. Por exemplo, os dois registros abaixo serão considerados válidos se aparecerem neste arquivo:

   ```
   163;725886;6743;112.5
   200;725156;398;15.0
   ```

5. Arquivo com dados de rotas - este arquivo contém uma relação de todas as rotas existentes no banco de dados da ESCELSA, com algumas informações relativas às rotas. O arquivo deve conter as seguintes colunas:

   - dw_rota - chave primária da rota no banco de dados da ESCELSA.
   - total_consumo - soma do consumo de todos os clientes da rota no último mês.
   - total_clientes - total de consumidores daquela rota no último mês.

   Este arquivo não deve ter ponto e vírgula no final de cada registro. Por exemplo, os dois registros abaixo serão considerados válidos se aparecerem neste arquivo:

   ```
   288;45;1
   2260;34;3
   ```

6. Arquivos com dados de atividades - este arquivo contém uma relação de todas as atividades existentes no banco de dados da ESCELSA, com algumas informações relativas às atividades. O arquivo de conter as seguintes colunas:

   - dw_atividade - chave primária da atividade no banco de dados da ESCELSA.
   - total_consumo - soma do consumo de todos os clientes da atividade no último mês.
   - total_consumidores - total de consumidores daquela atividade no último mês.

   Este arquivo não deve ter ponto e vírgula no final de cada registro. Por exemplo, os dois registros abaixo serão considerados válidos se aparecerem neste arquivo:

```
34;2911;4
35;440;4
```

7. Arquivo com dados de curva típicas - este arquivo contém as curvas típicas de todos os consumidores da ESCELSA. A curva típica nada mais é do que a média de consumo mês a mês de todos os consumidores, agrupados segundo algum critério. Atualmente estão sendo gerados dois tipos distintos de curvas típicas: curva típica por setor econômico e curva típica por município e classe. Conforme já exibido, a interface de conversão possui opção para indicação de qual curva típica está sendo usada. Os arquivos dos dois tipos de curvas típicas são diferentes, portanto se o tipo de curva não for corretamente marcado na interface de conversão o sistema não conseguirá converter os arquivos.

   A seguir os dois tipos de curvas são mostrados.

   Curvas típicas por setor econômico - contém a média de consumo de todos os consumidores agrupados por setor econômico e classe de serviço. O arquivo deve conter as seguintes colunas:

   - dw_setor_economico - chave de identificação do setor econômico
   - dw_class_serv - chave de identificação da classe de serviço
   - ano_mes - mês de referência
   - saldo - média de consumo de todos os consumidores no mês (agrupados por setor econômico e classe de serviço)

   Este arquivo não deve ter ponto e vírgula no final de cada registro. Por exemplo, os dois registros abaixo serão considerados válidos se aparecerem neste arquivo:

   ```
   4;1;200201;323.8
   2;1;200202;330.6
   ```

   Curvas típicas por município e classe - contém a média de consumo de todos os consumidores agrupados por município, classe, sub-classe e classe de serviço. O arquivo deve conter as seguintes colunas:

   - dw_municipio_com - chave de identificação do município
   - cod_classe - código da classe
   - cod_sub_classe - código da sub-classe
   - dw_class_serv - chave de identificação da classe de serviço
   - ano_mes - mês de referência
   - saldo - média de consumo de todos os consumidores no mês (agrupados por município, classe, sub-classe e classe de serviço)

   Este arquivo não deve ter ponto e vírgula no final de cada registro. Por exemplo, os dois registros abaixo serão considerados válidos se aparecerem neste arquivo:

   ```
   54;2;203;8;200307;2522
   69;2;205;8;200307;13817
   ```

Apêndice B

Formato do arquivo CSV gerado pelo conversor

A interface de conversão dos arquivos de saída do Datawarehouse gera como produto um arquivo CSV (disponível dentro do diretório “conversao” do sistema) com as colunas descritas a seguir. Todas as configurações de base de dados pré-definidas, disponíveis no diretório “conf/modelos” e descritas na seção 5.4.4, estão preparadas para receber os arquivos neste formato.

Vale destacar que, caso seja necessário alterar o formato da conversão, basta alterar os procedimentos descritos no apêndice C e reconfigurar a base de dados usando a interface de configuração de base de dados do sistema.

- dw_uc - contém a chave de identificação dos consumidores no Datawarehouse da ESCELSA. No arquivo CSV de saída do conversor, esta coluna é o identificador único dos exemplos.
- maior_mes_consumo - guarda o último mês no qual o consumidor em questão possui consumo diferente de nulo. Esta coluna tem sido usada para criação das colunas de consumo, conforme descrito a seguir.
- cod_classe - código que representa a classe a qual o consumidor pertence (residencial, comercial, industrial etc.).
- cod_sub_classe - código que representa a sub-classe a qual o consumidor pertence. Esta coluna é uma sub-divisão da classificação da coluna anterior.
- dw_class_serv - código que indica a classe de serviço do consumidor (monofásico, bifásico, trifásico ou primário).
- cod_tarifa - código da tarifa do consumidor. Este código é uma subdivisão das três colunas anteriores (classe, sub-classe e classe de serviço).
- cod_setor_econ - guarda o código do setor econômico do consumidor. O código do setor econômico é um agrupamento das atividades, também disponíveis na base.
- dw_setor_economico - guarda a chave interna do Datawarehouse de identificação do setor econômico. Necessária para geração das curvas típicas.
- cod_atividade - código da atividade do consumidor (padaria, indústria de mármore etc.).

- dw_gpo_fornec - código do grupo de fornecimento do consumidor (A, B etc.).
- dw_sub_gpo_fornec - código do sub-grupo de fornecimento do consumidor (A1, A2 etc.)
- dw_cate_cliente - código da categoria do cliente.
- dw_municipio_com - código do município da unidade consumidora do cliente.
- dsc_municipio - nome do município da unidade consumidora do cliente.
- dw_bairro_pt - código do bairro do consumidor.
- nome_bairro - nome do bairro do consumidor.
- dw_rota - código da rota de leitura do consumidor. Esta coluna é um indicador de vizinhança entre os exemplos da base.
- media_rota - média de consumo de todos os consumidores da rota do consumidor. A média é calculada sobre o último mês de consumo dos consumidores.
- dw_alimentador - código interno do alimentador ao qual o consumidor está ligado.
- nume_trafo - código do transformador ao qual o consumidor está ligado.
- potencia - potência nominal do transformador ao qual o consumidor está ligado.
- fator_carga - fator de carga do transformador ao qual o consumidor está ligado. O fator de carga é uma relação entre a potência nominal e o total de consumo de um mês de um transformador.
- media_atividade - média de consumo da atividade do consumidor no último mês.
- situacao - indica se o consumidor está ou não ligado no mês corrente.
- total_nulos - total de nulos na série de consumo do consumidor.
- curva - série temporal com a curva típica do consumidor. Na base de dados esta coluna é representada por 12 valores, nomeados como curva12, curva11, curva10, ..., curva01. O valor mais antigo (registro cujo mês é o mais distante do mês corrente) é chamado de curva12. O mais recente, curva01. A curva típica do consumidor é a média de consumo de todos os consumires do mesmo tipo que o consumidor corrente, mês a mês. Existem duas formas de determinar qual o tipo do consumidor: por setor econômico (dw_setor_economico) e classe de serviço ou pelo conjunto formado por município, classe, sub-classe e classe de serviço. Em cada caso mudará a forma que os consumidores serão agrupados para o cálculo do consumo típico. Os 12 valores da curva típica se referirão aos mesmos meses que compõe a série de consumo.
- calculo - série de consumo do consumidor. Assim como no caso da curva típica, é composta por 12 valores nomeados com uma marca de numeração no nome da coluna. Colunas com números maiores se referem a meses mais distantes. A série de consumo é retirada do histórico de consumo do consumidor de duas formas distintas. No caso da conversão iniciada a partir do Módulo de Aprendizagem, o sistema gerará a série de consumo para cada consumidor com os valores dos consumos imediatamente anteriores à data de inspeção. No

caso da conversão a partir do Módulo de Consultas, será usada como data de referência para recuperação do consumo o valor da coluna maior_mes_consumo.

- dw_motivo_ss - contém um código indicador do motivo de inspeções do consumidor. Caso o consumidor não tenha sido inspecionado (fato comum na preparação da base de consulta) o valor não estará preenchido.
- resultado - contém o resultado da inspeção. Os valores possíveis são: zero, para indicar que o consumidor não é fraudador; um, para indicar que o consumidor é fraudador. No caso da conversão para o Módulo de Consultas, esta coluna não conterá nenhum valor (nulo).

Cada linha do arquivo CSV criado pelo conversão conterá um exemplo distinto.

Apêndice C

Notas sobre os procedimentos de conversão

Nesta seção alguns detalhes do procedimento de conversão são descritos. As informações presentes nesta seção são destinadas às pessoas que necessitarem alterar os passos de conversão do sistema e não serão importantes na utilização diária do sistema.

A interface de conversão do sistema permite que o usuário selecione sete arquivos para conversão. Durante o processo de conversão, o sistema MIP executa dois arquivos que contêm programas em linguagem SQL. Os programas SQL são executados diretamente do servidor de Banco de Dados PostgreSQL.

Inicialmente existem cinco arquivo com fontes SQL no diretório "conf" da instalação do sistema MIP: um que é executado todas as vezes que uma nova conversão for iniciada (script_geral.sql) e quatro que são executados de acordo com as escolhas do usuário. São eles:

- script_curva1_treinamento.sql - executando quando o usuário aciona a interface de conversão a partir do Módulo de Aprendizagem e seleciona a curva típica por município e classe.
- script_curva1_consulta.sql - executando quando o usuário aciona a interface de conversão a partir do Módulo de Consultas e seleciona a curva típica por município e classe.
- script_curva2_treinamento.sql - executando quando o usuário aciona a interface de conversão a partir do Módulo de Aprendizagem e seleciona a curva típica por setor econômico.
- script_curva2_consulta.sql - executando quando o usuário aciona a interface de conversão a partir do Módulo de Consultas e seleciona a curva típica por setor econômico.

A principal função destes programas é converter os sete arquivos de entrada do Datawarehouse (no formato descrito na seção A) em uma arquivo CSV (no formato descrito na seção B).

A principal restrição deste método de conversão é a importação dos arquivos de dados para o servidor de bando de dados. Ainda não é suportado que o gerenciador de banco de dados esteja executando em computador diferente do MIP. Maiores informações sobre os procedimentos de conversão podem ser obtidas na documentação do código do programa.

Apêndice D

Glossário

1. Classe: é um atributo especial do exemplo que irá conter a informação de interesse, isto é, a informação que se deseja aprender e fazer previsões a respeito (no caso do MIP, se o consumidor é ou não é candidato à inspeção).
2. Técnica de classificação: o objetivo de uma técnica de classificação é gerar um bom classificador a partir de um conjunto de registros que possuem a classe conhecida.
3. Classificador: objeto gerado a partir dado um conjunto de registros e uma técnica de classificação. O objetivo do classificador é, dado um novo exemplo, ser capaz de predizer com a maior precisão possível a sua classe.
4. Estimador de erro: técnica de estimativa do desempenho futuro do classificador utilizando um conjunto de registros com classe conhecida.